Journal of Engineering and Technology for Industrial Applications

ISSN 2447-0228

February 2026

Volume 12/ No 57



Editor-in-Chief: J. C. Leite

www.itegam-jetia.org

ITEGAM - JETIA ISSN 2447-0228 Online
Journal of Engineering and Technology for Industrial Applications (JETIA)

O **ITEGAM-JETIA: Journal of Engineering and Technology for Industrial Applications** is a publication of the Galileo Institute of Technology and Education of the Amazon (ITEGAM), located in the city of Manaus since 2008. JETIA publishes original scientific articles covering all aspects of engineering. Our goal is the dissemination of research original, useful and relevant presenting new knowledge on theoretical or practical aspects of methodologies and methods used in engineering or leading to improvements in professional practice. All the conclusions presented in the articles It should be state-of-the-art and supported by current rigorous analysis and balanced assessment. Public magazine scientific and technological research articles, review articles and case studies.
JETIA will address topics from the following areas of knowledge: Mechanical Engineering, Civil Engineering, Materials and Mineralogy, Geosciences, Environment, Information and Decision Systems, Processes and Energy, Electrical and Automation, Mechatronics, Biotechnology and other Engineering related areas.

Publication Information:
ITEGAM-JETIA (ISSN 2447-0228), (online) is published by Galileo Institute of Technology and Education of the Amazon on a every two months (February, April, June, August, October and December).

Contact information:

Web page: www.itegam-jetia.org
Email: editor@itegam-jetia.org
Galileo Institute of Technology and Education of the Amazon (ITEGAM).

Joaquim Nabuco Avenue, No. 1950. Center. Manaus, Amazonas. Brazil.
Zip Code: 69020-031. Phone: (92) 3584-6145.

Copyright 2014. Galileo Institute of Technology and Education of the Amazon (ITEGAM)
The total or partial reproduction of texts related to articles is allowed, only if the source is properly cited. The concepts and opinions expressed in the articles are the sole responsibility of the authors.

Previous Notice
All statements, methods, instructions and ideas are the sole responsibility of the authors and do not necessarily represent the view of ITEGAM -JETIA. The publisher is not responsible for any damage and / or damage to the use of the contents of this journal. The concepts and opinions expressed in the articles are the sole responsibility of the authors.

Directory

Members of the ITEGAM Editorial Center - Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA) of the Galileo Institute of Technology and Education of the Amazon (ITEGAM). Manaus-Amazonas, Brazil.

Jandecy Cabral Leite, CEO and Editorial Editor-in-Chief
Ivan Leandro Rodriguez Rico, Editorial Assistant
Jorge de Almeida Brito Junior, Editorial Assistant

ITEGAM - JETIA ISSN 2447-0228 Online
Journal of Engineering and Technology for Industrial Applications (JETIA)

ITEGAM-JETIA. v.12, n.57. February of 2026. Manaus - Amazonas, Brazil. ISSN 2447-0228 (ONLINE)
https://www.itegam-jetia.org

SUMMARY

INTELLIGENT CONTROL STRATEGIES FOR DFIG-BASED WIND ENERGY SYSTEMS: A COMPARATIVE ANALYSIS OF TYPE-1 FUZZY, INTERVAL TYPE-2 FUZZY, AND ANN APPROACHES **01**
Kouadria Selman, Kouadria Mohamed Abdeldjabbar, Berkouk El Madjid, Rizoug Nassim

MULTI-OBJECTIVE OPTIMIZATION OF CMOS FOLDED-CASCODE OTAS USING DIVERSE EVOLUTIONARY ALGORITHMS **16**
Alpesh M. Patel, Pankaj P. Prajapati, Anilkumar J. Kshatriya, Kirit V. Patel, Ghanshyamkumar Sah, Chintan Dave, Kinjal R. Sheth, Devendra H. Patel

COMPARATIVE STUDY OF HMM AND BPNN IN DETECTING CORONA DISCHARGE ON 20 KV CUBICLE BASED ON VOLTAGE AND SOUND **24**
Christiono Christiono, Miftahul Fikri, Syamsir Abduh

DESIGN AN INTELLIGENT CONTROLLER FOR LOAD BALANCING IN WIND ENERGY SYSTEMS CONNECTED TO THE GRID **35**
Zainab Jamal Mohammed, Mafaz Yahya Damarah

DARK PATTERNS IN DIGITAL DESIGN: A REVIEW AND ASSESSMENT VIA RISK, TRUST AND DIVERSITY SCORES **43**
Preeti Marwaha, Shalu Mahajan, Ishitva Joshi, Kriti Mishra, Ananya Srivastava, Priyanka Sharma, Arunita Chaukiyal

MODELING AND SIMULATION OF TEMPERATURE EFFECTS ON INGAP/INGAAS/GE TRIPLE-JUNCTION SOLAR CELLS WITH PYAMS **59**
Dhiabi Fathi, Okba Benelmir

PREDICTING CARDIOVASCULAR DISEASES THROUGH INTEGRATED DEEP LEARNING FEATURES **69**
Islam Daoud Suliman, D. Vasumathi

AN ENHANCED MULTIMODAL APPROACH FOR SENTIMENT ANALYSIS USING DEEP LEARNING TECHNIQUES **77**
Vinitha V, S.K.Manju bargavi

OPTIMIZING OVERLAP IN TREE-BASED INDEXING STRUCTURES FOR ENHANCED K-NN SEARCH EFFICIENCY **86**
Ala Eddine Benrazek, Ibtissem Kemouguette, Zineddine Kouahla, Brahim Farou, Hamid Seridi, Pr.

MACHINE LEARNING FOR CORONARY HEART DISEASE RISK PREDICTION: A CONVENTIONAL AND ENSEMBLE CLASSIFICATION MODEL APPROACH **104**
Raghavendran Ch V, Chandra Mouli Kathi, Bhargavi Latha S

CARDIOVASCULAR RISK PREDICTION USING AI **115**
Mehek Mohd Meraj Qureshi, Hariram Chavan

DIFFERENTIABLEHMM: NEURAL DIFFERENTIABLE HIDDEN MARKOV MODEL **124**
Arbia Boudaoud, Salheddine Kabou

ENHANCING RENEWABLE ENERGY RELIABILITY: A STUDY ON HYBRID SOLAR-WIND SYSTEMS FOR EFFICIENT POWER GENERATION **133**
Omar Gheyath Ibrahim, Mohammed Tariq Majeed, Ghufran Saad Mohammed, Israa Khalouq Saaed

ITEGAM - JETIA
ISSN 2447-0228 Online
Journal of Engineering and Technology for Industrial Applications (JETIA)

FAULT PROTECTION ANALYSIS OF A DC MICROGRID USING VSC-BASED MODULAR MULTILEVEL CONVERTER (MMC) **147**
Kanna Sravanthi, Ayya Chaithanya

EXPERIMENTAL METHODS AND A/B TESTING IN FINTECH: APPLIED GUIDE IN PYTHON **156**
Alexander Fernando Haro Sarango, Julian Coronel Reyes, Johnny Javier Triviño Sanchez, Ibeth Molina-Arcos, Gerardo Villarreal-Terán, Gladys Elizabeth Proaño Altamirano

EFFECT OF BATTERED PILES ON THE RESPONSE OF PILED RAFT AND PILE GROUP FOUNDATIONS A NUMERICAL STUDY **167**
Toudert Bougrires, Sadok Benmebarek, Naïma Benmebarek

INFLUENCE OF DEFECTS ON THE MECHANICAL PROPERTIES OF METALLIC COATINGS OBTAINED BY THERMAL SPRAYING: A REVIEW **176**
Uirá Rodrigues da Silveira Montani, Marilia Garcia Diniz, Paolo Bettini, Hector Reynaldo Meneses Costa

PREDICTING CUSTOMER SATISFACTION THROUGH LAYERED NLP TECHNIQUES: TRADITIONAL MEETS DEEP LEARNING **188**
Mahdjoub Mohammed Bachir, laallam Fatima Zohra, Mezati Messaoud

ANALYSIS OF REACHING LAW EFFECTIVENESS IN SLIDING MODE CONTROL OF THE UPFC **201**
Boukhalfa Mohamed, Mohammed Elaguab, Abdelkrim Bouanane, Badreddine Naas, Abdelkader Azzeddine Bengharbi, Mohamed R. Bengourina

A BIG DATA BASED EMOTION DETECTION FRAMEWORK FOR SOCIAL MEDIA USING APACHE SPARK. **211**
Wafa Saadi, Fatima Zohra Laallam, Messaoud Mezati

HYBRID GPT AND NEURAL MODELS FOR PERSONALIZED E-COMMERCE: A NOVEL FRAMEWORK FOR ADAPTIVE AND TRANSPARENT PRODUCT RECOMMENDATIONS **222**
Pavan Gunda, Thirupathi Rao Komati

INVESTIGATION OF THE EFFECT OF NANOPARTICLES ON ENGINE PERFORMANCE AND EMISSION CHARACTERISTICS IN DIESEL ENGINES **236**
Sathiyaraj S, Imthiyas .A, Kumaran .P, Silamburajan .M, Rengaraju .G

PERFORMANCE AND EMISSION ENHANCEMENT OF A CRDI VCR ENGINE FUELED WITH TIO_2-INFUSED TOMATO SEED BIODIESEL USING RSM OPTIMIZATION **244**
Kumaran P, Sathiyaraj S, Vijayakumar K, Albinraj I, Esakki mani K

IMPLEMENTATION AND COMPARATIVE ANALYSIS OF A FUZZY LOGIC CONTROLLER FOR INDUCTION MOTOR VECTOR CONTROL **259**
Salim Mahdab, Mohamed Fekir

ENHANCED ECONOMIC DISPATCH OPTIMIZATION USING MOMENTUM IN THE ORCA PREDATION ALGORITHM **265**
Vivi AIda Fitria, Arif Nur Afandi, Aripriharta null, Azwar Riza Habibi

OPTIMIZATION OF CONVOLUTIONAL NEURAL NETWORKS THROUGH SOFT VOTING ENSEMBLE AND FINE-TUNED APPROACH TO IMAGE CLASSIFICATION **277**
Johan Wijaya Kusuma, Julian Supardi

THERMAL EVALUATION OF OPTIMAL MOLDING MATERIAL SELECTION THROUGH ANSYS SIMULATION **287**
Rama Dani Eka Putra, Agus Nuramal, Hendri Hestiawann

ITEGAM - JETIA
ISSN 2447-0228 Online
Journal of Engineering and Technology for Industrial Applications (JETIA)

DESIGN, FABRICATION, AND EVALUATION OF A LOW-COST AND MOBILE FISH SMOKEHOUSE FOR SMOKED SOFT-BONED MILKFISH **296**
Ervin Fernandez Papa, Jesus Paolo B. Cullado, Quennie M. Bilugan

SYNTHETIC DATA-DRIVEN ANALYSIS OF ELECTRIC VEHICLE MANUFACTURING IN INDIA USING MACHINE LEARNING **305**
Pasala Gopi, Tharun Korivi, Vinod Kumar Reddy P, Sri Ayodhya Raju K, Sathish Kumar A

AN ARTIFICIAL NEURAL NETWORK FRAMEWORK FOR AUTOMATED FAULT DIAGNOSIS IN INDUSTRIAL DC-DC CONVERTERS **315**
Naas Djeddaoui

ENHANCING SWELLING PRESSURE PREDICTION THROUGH INTEGRATED MACHINE LEARNING METHODS **323**
Laid Lekouara, Fatima Zohra Tebbi, Ouassila Bahloul, Rachid Rabehi

PERFORMANCE ASSESSMENT OF TEXT SIMILARITY ALGORITHMS THROUGH CLASSIFICATION METRICS FOR BSIT GRADUATE JOB CONCORDANCE **332**
Mark Gil Toribio Gañgan, Thelma D. Palaoag

COMPARATIVE STUDY OF ENERGY MANAGEMENT STRATEGIES FOR HYBRID ELECTRIC VEHICLE WITH HYBRID ENERGY STORAGE SYSTEM **340**
Sunka Divya, Danduprolu Kiran Kumar

EVALUATING THE MEAN TIME BETWEEN FAILURE (MTBF) OF CRITICAL BUILDING SYSTEMS IN STUDENTS' HOSTEL BUILDINGS WITHIN LAGOS METROPOLIS **348**
Olajide Julius Faremi, Damilola Abdurazaq Lukman, Dele Roger Simeon

FORMABILITY AND SIZE EFFECTS IN MICROFORMING: A REVIEW OF MATERIALS, METHODS, AND APPLICATIONS **366**
Gopi Durgam, Kishor K Dhande

AN EVALUATION OF DIRECT AND INDIRECT STRATEGIES FOR HIERARCHICAL NET-LOAD FORECASTING **375**
Oyebola Olasupo, Temitope Raphael Ayodele, Ayodeji Samson Olatunji Ogunjuyigbe, Prof.

A SELF–EVALUATION MATURITY ASSESSMENT MODEL FOR LEAN SIX SIGMA IMPLEMENTATION IN SMES **389**
Sushil Shyamacharan Mishra, Ravi Terkar

OPTIMIZATION OF LASER CUTTING MODES FOR PARTS MADE FROM SHEET STEEL 12X18H10T **400**
Dmitriy Podashev, Anton Kisel', Ivan Nikolaev

OPTIMAL CONVERSION TOPOLOGIES FOR GRID-CONNECTED PV SYSTEMS: A COMPARATIVE STUDY OF SINGLE-STAGE AND TWO-STAGE CONFIGURATIONS **408**
Abdelmoumen Ghilani, Amel Terki, Ahmed Marouane Ghodbane, Zakaria Alili, Oussama Belaroussi

EXPERIMENTAL INVESTIGATION ON MICROSTRUCTURE AND MECHANICAL PROPERTIES OF SINTERED 316 STAINLESS STEEL SPECIMEN FOR BIO IMPLANTS **416**
Imthiyas A, Bubesh Kumar D, Sangeetha krishnamoorthi S, Vasanth kumar J, Dhanush kumar R

ITEGAM - JETIA
ISSN 2447-0228 Online
Journal of Engineering and Technology for Industrial Applications (JETIA)

AI APPLICATIONS FOR DRONE OF THINGS IN THE POWER SECTOR **428**
Manjunatha H M, Madan H T, Poshitha B, Swapna H, Aruna B, Pramodh H K, Lokesh Mallaiah

ERGONOMICS THERMAL ENGINEERING APPROACH IN DESIGNING TRADITIONAL DRYER WITH RICE HUSK ENERGY FOR SMALL-SCALE RENGGINANG PRODUCTION **448**
I Gede Bawa Susana, Ida Bagus Alit

APPLICATION OF THE ANALYTIC HIERARCHY PROCESS (AHP) METHOD FOR CHOOSING A PORTABLE RETINOGRAPH IN OPHTHALMOLOGICAL CLINICAL PRACTICE **457**
Marcelo Maia do Nascimento, Railma Lima de Paula

A SYSTEM FOR DETECTING NON-CONFORMITIES IN INDUSTRIAL MANUFACTURING PROCESSES IN THE AUTOMOTIVE SECTOR USING FUZZY LOGIC **465**
Roberto Ferreira de Lima, Helder Kiyoshi Miyagawa, Rivanildo Duarte Almeida

INTELLIGENT MODEL FOR PRODUCTIVE PLANNING OF ELECTRONIC METERS WITH INDUSTRY 4.0 TECHNOLOGICAL RESOURCES **483**
Jandecy Cabral Leite, Marcelo Maia do Nascimento, Marivan Silva Gomes, Alan Ferreira Pinheiro Tavares, André Barbosa de Lima, Matheus Pereira de Souza, Daniel Gomes Gulart, Marcos Castro de Menezes

AUTOMATED CONTROL AND PLANNING SYSTEM FOR THE PRODUCTION OF ELECTRONIC METERS WITH ARTIFICIAL INTELLIGENCE **496**
Jandecy Cabral Leite, Marcelo Maia do Nascimento, Marivan Silva Gomes, Alan Ferreira Pinheiro Tavares, André Barbosa de Lima, Matheus Pereira de Souza, Daniel Gomes Gulart, Marcos Castro de Menezes

QUANTIFYING THE CREDIBILITY OF E-LEARNING SYSTEMS USING THE BERT MODEL **519**
Piyush Singh, Rohan Pachisiya, Sagar Nehra, Vibha Gaur

ANALYSIS OF S.S.43100 BY TURNING OF TI16ZR PVD COATED TOOL INSERTS AND OPTIMIZATION OF SURFACE ROUGHNESS USING TAGUCHI METHOD **529**
Nishadevi N. jadeja, Sanjay .H.Zala, Hirendra .G.Vyas

INTELLIGENT GAME-THEORETIC DEEP REINFORCEMENT LEARNING CACHING FOR NAMED DATA NETWORKING: A UNIFIED OPTIMIZATION FRAMEWORK **535**
Djamal Seghier, Maaskri Moustafa, Goismi Mohamed, Debbab Mohamed

INTEGRATED MCDM MODEL FOR EVALUATION OF TWO-FILE PREQUALIFICATION TENDER AT PROCUREMENT BUREAU IN INDONESIA **551**
Citra Amalia, Afiza Ismail, Emma Nuraihan Mior Ibrahim, Shuhaida Mohamed Shuhidan, Norzehan Sakamat

A THREE-TIER INTEGRATED QUEUE MANAGEMENT SYSTEM: OPTIMIZED SERVICE DELIVERY FOR LAND TRANSPORTATION CLIENTS **559**
Ines Gimota Falcon, Kenneth Jay L. Dugaria, Danna C. Remojo

STRUCTURAL OPTIMIZATION USING GREY WOLF OPTIMIZER AND ITS VARIANTS **567**
Bhavik Dipakbhai Upadhyay, Ms. Sonal T. Dave, Vishal J. Pandya, Jignesh G. Parmar, Hasmukh P. Koringa, Miral J. Patel

AREA AND POWER EFFICIENT SPARSE AWARE ARCHITECTURE FOR POINTWISE CONVOLUTION UNIT USING VARIABLE DENSITY BOUND BLOCK **574**
Golajapu Jyothsna, Narayanam Balaji

ITEGAM - JETIA
ISSN 2447-0228 Online
Journal of Engineering and Technology for Industrial Applications (JETIA)

INTEGRATION OF MOST AND TOPSIS METHODS TO SUPPORT BUSINESS DEVELOPMENT STRATEGY FORMULATION IN THE STEEL PIPE INDUSTRY OF SURABAYA, INDONESIA **581**
Mohamad Ferdaus Noor Aulady, Rachmat Yustiawan Hadi

PERFORMANCE ANALYSIS OF EMOTION RECOGNITION USING YOLO **588**
Varsha Patil, Varun Ashish Chitale, Vijaya Pawar, Megha Yannwar

IOT-BASED MULTI-SENSOR DATA FUSION FOR PRECISION CROP YIELD OPTIMIZATION USING ARDUINO, ESP32, AND WEBCAM INTEGRATION **602**
Chandraiahgari Dinesh Reddy, Danduprolu Kiran Kumar

COMPREHENSIVE ENERGY AUDIT AND EFFICIENCY ASSESSMENT OF A UNIVERSITY CAMPUS: A CASE STUDY AT THE UNIVERSITY OF EASTERN PHILIPPINES **612**
Bless Gordo Ampuan, Rem Laodeño, Merewina Llanie Tapong

REVIEW OF EMERGING AI TECHNIQUES FOR GREEN HYDROGEN IN SUSTAINABLE ENERGY **620**
Hiba abdulkareem Saleh, Rana Khalid Sabri, Omar Ibrahim Alsaif

ENHANCED CONTROL OF DFIG-BASED WIND TURBINE USING HYBRID FUZZY SUPER-TWISTING SLIDING MODE CONTROL FOR IMPROVED POWER QUALITY AND ROBUSTNESS **628**
Anis Feddaoui, Lotfi Farah, Abdelouahab Benretem, Mohammed Abdeldjalil Djehaf, Youcef Djeriri

A NOVEL ADAPTIVE ENSEMBLE INTELLIGENCE FRAMEWORK FOR STRENGTHENING CLOUD SECURITY AGAINST DDOS THREATS FOR ACCURATE DETECTION AND MITIGATION **637**
Abida T, M. Shanmugapriya

A HYBRID FUZZY-GRADIENT MPPT METHOD FOR ENHANCING PHOTOVOLTAIC EFFICIENCY **648**
Mahdi Bennoui, Ali Boukhari, Chouaib Labiod, Redha Meneceur, Abdelmoumene Hechifa, Ali Bebboukha

TO3: A COMPACT AND EFFICIENT ENCODING SCHEME FOR SKEWED SLICING FLOORPLANS IN VLSI DESIGN **661**
Biswojit Nayak, Mrutynjaya Panda

PREDICTING STUDENTS' CONCENTRATION IN COGNITIVE ACTIVITIES USING EEG AND DEEP LEARNING TECHNIQUES **671**
Harsha Gaikwad, Yogesh Patil, Snehal Mali, Sanil Gandhi, Manjushree Laddha, Arvind Kiwelekar

BREAST CANCER DETECTION USING LATE FUSION APPROACH ON MULTIMODAL DATA WITH DEEP LEARNING **683**
Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Az-Zahra, Fitri Utaminingrum

USE OF DUNE SAND TREATED WITH HYDRAULIC BINDERS IN ROAD FOUNDATION LAYERS **696**
Ali Smaida, Dif Fodil, Khalil Latreche, Smail Haddadi, Hayet Cherfa

DC-BUS CAPACITOR DISCHARGE CONTROL STRATEGY FOR PMSM DRIVES SYSTEM: ELECTRIC VEHICLES **705**
Chevuru Mahima, Chinthapudi Chengaiah

PERFORMANCE EVALUATION OF JSON LIBRARIES IN KOTLIN SERVER ENVIRONMENTS: JACKSON, GSON, AND KOTLINX SERIALIZATION **718**
Achmad Arwan, Julia Nur Fajrina, Eriq Muhammad Adam Jonemaro

ITEGAM - JETIA
ISSN 2447-0228 Online
Journal of Engineering and Technology for Industrial Applications (JETIA)

TOWARD ETHICAL AND INTELLIGENT SENTIMENT ANALYSIS IN TOURISM: FUTURE DIRECTIONS FOR THE NEXT TWO DECADES **728**
Hardian Oktavianto, Ika Safitri Windiarti, Bagus Setya Rintyarna

ENHANCING SUSTAINABILITY THROUGH WASTE HEAT UTILIZATION: A STUDY OF BRAZING FURNACE HEAT RECOVERY **738**
Sundaramahalingam Subramaniam

TRENDS, COLLABORATION NETWORKS, AND RESEARCH IMPACT IN THERMAL MANAGEMENT OF ELECTRONIC DEVICES USING PASSIVE COOLING TECHNIQUE **747**
Nilesh Gaikwad, Narendra Deore, Mahadev Madgule

EXPERIMENTAL STUDY ON NR–SBR RUBBER COMPOSITE WITH CARBON BLACK FILLER FOR INDUSTRIAL TRACK PAD APPLICATIONS **763**
Farit Hendro Wibowo, Sujito Sujito, Muhammad Rodhi Faiz

RENEWABLE ENERGY AND RESOURCE RECOVERY SYSTEMS USING AI TOOLS **774**
Gunasekaran Priya, Murugesan Kamaraj, Aravind Jeyaseelan

A LIGHTWEIGHT FEDERATED PREDICTION APPROACH FOR URBAN VRU MOVEMENT UNDERSTANDING IN AUTONOMOUS DRIVING **787**
Lakshmi Narayana I, T M N Vamsi

VOICE DISORDER DETECTION: HYBRID GTCC-MFCC FUSION WITH SOURCE-BASED FEATURES OPTIMIZED FOR THE MALE SUB-COHORt **798**
Aboubakr Missaoui, Fatima Chouireb, Boubakeur Latreche, Abdelkerim Souahlia, Messaoud Linani, Abdelaziz Rabhi

SUSTAINABLE STABILIZATION OF LOESS SOILS USING MARBLE POWDER: IMPROVEMENT OF MECHANICAL PROPERTIES **807**
Mohamed Dacher, Samir Benmoussa, Ouassila Bahloul, Mohamed Salah Laouar

ENHANCED FACE RECOGNITION SYSTEM INTEGRATING GAN-AUGMENTED CNN FEATURES WITH ICA AND VISION TRANSFORMER **820**
Sangeetha Karunakaran, A. Akila

RISK-BASED PERFORMANCE OPTIMIZATION OF MOZZARELLA CHEESE PROCESSING USING EXPERT JUDGEMENT – PREFERENCE SELECTION INDEX (PSI) MODEl **832**
Vaishali Talkit, Anand Bewoor, Amitkumar Shelar

SCIENTIFIC RESEARCH AS A KEY ELEMENT OF INSTITUTIONAL TRANSFORMATION IN THE PERUVIAN ARMY: A SYSTEMATIC REVIEW **846**
Michael Joaquin Herrera Contreras, Aaron Aguirre Venturi, Ada Gallegos

EARLY ALZHEIMER'S DISEASE DETECTION BY USING RESNET FOR EXTRACTION OF FEATURES WITH SVM TUNING FOR CLASSIFICATION **856**
Rama Lakshmi Boyapati, Ch Rama Devi, Malathi J anapati, Mirtipati Satish Kumar, Anjani Y, Pavan Gunda

ACCIDENT PREVENTION IN THE DAIRY SECTOR: EVALUATION OF A SAFETY PROGRAM **869**
Víctor Aldair Custodio Ríos, Luis Sleither Huaman Tantacure, Mirtha Yvis Santisteban Salazar

PREDICTING POSTPARTUM DEPRESSION RISK USING CROSS VECTOR SPIDER SWARM INTELLIGENCE AND HYPERNET GATED MULTI-PERCEPTRON NEURAL NETWORK **876**
Jomila Ramesh, V. VishwaPriya

ITEGAM - JETIA
ISSN 2447-0228 Online
Journal of Engineering and Technology for Industrial Applications (JETIA)

EVALUATION OF THE PERFORMANCE OF COMPUTATIONAL MODELS FOR THE ANALYSIS OF THE PROTECTION OF MIRROR COATINGS WITH DATA OBTAINED USING ELECTROCHEMICAL IMPEDANCE SPECTROSCOPY **888**
Juan Carlos Gonzalez Islas, Ernesto Bolaños-Rodriguez, Asdrúbal López-Chau, Gildardo Godinez-Garrido, Aldo Márquez-Grajales, Fernando Salas-Martínez

ESTIMATING THE SERVICE LIFE OF CONCRETE REINFORCED WITH ADVANCED MATERIALS: INTEGRATION OF IOT SENSORS AND DIGITAL TWINS **896**
Neyser Alcides Pérez, Socrates Pedro Muñoz, Victor Manuel Valdiviezo Sir

SYSTEMATIC REVIEW OF THE USE OF FLY ASH IN GEOPOLYMER CONCRETE: IMPACT ON ITS PHYSICAL, MECHANICAL, AND MICROSTRUCTURAL PROPERTIES **906**
Cristian Chuquipul Angulo, Socrates Pedro Muñoz Pérez, Victor Manuel Valdiviezo Sir, Mg. Victor Valdiviezo

MONITORING DIGITAL TWIN FRAMEWORK FOR CONTROLLED ENVIRONMENT AGRICULTURE USING DEEP LEARNING AND EDGE COMPUTING **918**
Sasikala S, Sita Devi Bharatula

RANDOM FOREST GUIDED FEATURE WEIGHTED KNN MODEL FOR NUCLEAR REACTOR EVENT IDENTIFICATION **930**
Arunprasath V, Santhosh T.V, Gopika Vinod

SMART PLS STRUCTURAL MODELLING OF AI-DRIVEN HR PRACTICES, ENTREPRENEURIAL SUCCESS AND EMPLOYER REPUTATION IN INDIAN FIRMS **941**
Nidhi Sonkar, Amit Kumar Pandey, Ratchagaraja Dhairiyasamy, Deekshant Varshney, Subhav Singh

A DATA-DRIVEN COMPETENCY DETECTION MODEL FOR IMPROVING WORKFORCE PLANNING AND DECISION-MAKING **951**
Lucrecia Llerena, Nancy Rodriguez, Jefferson Bustamante, Angel Torres

PERFORMANCE AND EMISSION OPTIMIZATION OF A DIESEL ENGINE USING KARANJA BIODIESEL–HYDROGEN DUAL-FUEL WITH AZADIRACHTA INDICA ANTIOXIDANT **964**
Kumaran P, Natarajan S, Sathiyaraj S, Subramanian T

DYNAMIC VIBRATION SUPPRESSION OF SMART COMPOSITE PLATES VIA LQR AND FINITE ELEMENT METHOD **976**
Mourad Chalane, Mohamed Latrache, Bachir Labiodh

A LIGHTWEIGHT PRIVACY-PRESERVING DATA STORAGE USING MODIFIED RING SIGNATURE SCHEME BASED ON ELLIPTIC CURVE CRYPTOGRAPHY **984**
Vasughi T.S, Muthulakshmi P

ENHANCING DISEASE DIAGNOSIS IN HEALTHCARE USING LSTM NETWORKS WITH OPTIMIZED FEATURE SELECTION: A COMPARATIVE RESEARCH ON HEART DISEASE, BREAST CANCER, AND LIVER DISEASE DATASETS **993**
Tina Sachdeva, Priyanka Sharma, Mehtab Alam

DEVELOPMENT OF A MODEL FOR ASSESSING THE SEISMIC RELIABILITY OF BRIDGES TO ENHANCE THEIR RESILIENCE **1009**
Mohammed Abdellaoui, Mohamed Badaoui, Said Koriga; Mahmoud Bensaibi

ADAPTIVE MULTI-STAGE CLOUD INTRUSION DEFENCE AND RECOVERY VIA HIERARCHICAL FEATURE OPTIMIZATION AND FEDERATED INTELLIGENCE SETS **1016**
Shubhangi Shailesh Shambharkar, Latesh G. Malik

ITEGAM - JETIA
ISSN 2447-0228 Online
Journal of Engineering and Technology for Industrial Applications (JETIA)

ENERGY CONSUMPTION PREDICTION IN 3D PRINTING USING ENHANCED HYBRID DEEP LEARNING MODELS **1030**
P. Manoj Kumar, M. Jayakrishna, M. Vijay

FEDERATED LEARNING FOR ENHANCING RELIABILITY AND SECURITY IN MEDICAL IMAGE ANALYSIS AGAINST ADVERSARIAL THREATS **1045**
Meenu Mohil, Akshita Mohil, Abhiraj Singh Mohil

HOSPITAL INFORMATION SYSTEM UTILIZATION MODERATES DIGITAL SELF-EFFICACY, PERSON–JOB FIT, AND WORK–LIFE BALANCE ON HEALTHCARE SERVICE PERFORMANCE **1054**
Paramita Adinda Putri, Adelia Usi Lauditta, Gracia Shinta S. Ugut

OPTIMAL ENERGY-EFFICIENT INTEGRATION OF WIND DISTRIBUTED GENERATORS IN A HYBRID AC-DC NETWORK USING A TUNICATE SWARM ALGORITHM **1063**
Mohamed Zellagui, Samir Settoul, Claude Ziad El-Bayeh

OF AHERO RICE SURVIVAL ANALYSIS WITH COMPETING RISKS: INSIGHTS FROM THE FINE AND GRAY MODEL AND COVARIATE ADJUSTMENTS **1073**
Anu Swedha Ananthan, Sumithra B, Dr, A.ZechariahJebakumar Arulanantham, Dr

ENHANCED SPAM DETECTION USING MUTUAL INFORMATION FEATURE SELECTION AND JAYA-TUNED TABTRANSFORMER **1084**
Mohamed Goismi, Mohamed Debbab, Moustafa Maaskri, Djamal Seghier

COMPUTATIONAL FLUID DYNAMICS APPLICATIONS IN VERTICAL AND INDOOR FARMING: A BIBLIOMETRIC REVIEW OF MICROCLIMATE AND VENTILATION STUDIES **1094**
Edwin Andres Villagran Munar, John Espitia, Jorge Flores-Velasquez, Cruz Ernesto Aguilar, Jader Rodriguez

AN APPROACH TO AUTOMATION IN SMART WASTE CLASSIFICATION BASED ON MACHINE LEARNING FOR EMBEDDED DEVICES **1110**
Pham Trung Kien

DYNAMIC COMBINED ECONOMIC EMISSION DISPATCH IN RENEWABLE ENERGY MICROGRIDS USING ZEBRA OPTIMIZATION ALGORITHM **1120**
Ram Prasad Kannemadugu, Y. Venkata Krishna Reddy, Katabathina Venkata Subbaiah

SECURE AND INTELLIGENT SDN ROUTING AND ANOMALY DETECTION USING XGBOOST FOR REAL-TIME IOT TRAFFIC OPTIMIZATION **1132**
Sohaib R. Awad, Qutaiba I. Ali, Amar I. Daood

RANDOMIZED DENDRITIC NEURON MODEL PRESERVING BIOLOGICALLY INSPIRED DENDRITIC GATING FOR EFFICIENT MEDICAL DIAGNOSTICS **1148**
Michael Angello Qadosy Riyadi, Adinda Mariasti Dewi

HIGH-ACCURACY BINARY CLASSIFICATION OF DIABETIC RETINOPATHY FROM FUNDUS IMAGES USING CNN ARCHITECTURE **1162**
Forhad Javed, Ashraful Islam, Mohammed Hasibul Hasan Chowdhury, Sayma Sultana, Nazmul Islam Frahim, Atikur Rahma

DIGITAL TRANSFORMATION OF THE PUBLIC FINANCE SYSTEM: CHALLENGES, INCLUSIVITY, AND PROSPECTS FOR SUSTAINABLE DEVELOPMENT **1170**
Oleksandr Bulavynets, Oleksandr Shashkevych, Sviatoslav Gerchakivskiy, Nataliia Gavkalova, Oleksandr Atamas

ITEGAM - JETIA
ISSN 2447-0228 Online
Journal of Engineering and Technology for Industrial Applications (JETIA)

FUZZY ADAPTIVE SPEED AND POSITION CONTROL FOR PERMANENT MAGNET SYNCHRONOUS MOTORS **1181**
Bentchikou Ibrahim, Tlemçani Abdelhalim, Nouri Hassan, Boudjema Fares, Boukhetala Djamel, Ould cherchali Noureddine, Fekir Mohamed, Mahdab Salim

PARTICLE SWARM OPTIMIZATION AND ENSEMBLE METHODS FOR ABSENTEEISM PREDICTION **1197**
Tabarek Alwan Tuib, Baydaa Hadi Saoudi, Yaqdhan Mahmood Hussein

META-LEARNING BASED STACKING OF CNN, SVM, AND RANDOM FOREST FOR MULTI-SPECIES CROP LEAF PATHOLOGY RECOGNITION **1207**
Guefrouchi Ryma, Amrouch Siham

THE FOUNDATION OF A DEFECT TRACKING MATRIX AND CORRECTIVE MEASURES FOR DEFECT REDUCTION IN THE CRANKSHAFT FORGING INDUSTRY: A PATH TOWARD SUSTAINABLE MANUFACTURING PRACTICES & DRIVING INNOVATION IN QUALITY CONTROL **1218**
Amitkumar Bharatkumar Solanki, Nisha Bora, Nilesh H. Pancholi, Nilesh Tank

ARTIFICIAL NEURAL NETWORK BASED PREDICTIVE PERFORMANCE ANALYSIS OF PHOTOVOLTAIC OUTPUT UNDER VARIOUS SOLAR RADIATION RANGES **1229**
Ravshanbek Rakhmatulaev, Rodney Tan, Aliev Rayimjon, Yu L. J

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 01-15. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2372

RESEARCH ARTICLE OPEN ACCESS

INTELLIGENT CONTROL STRATEGIES FOR DFIG-BASED WIND ENERGY SYSTEMS: A COMPARATIVE ANALYSIS OF TYPE-1 FUZZY, INTERVAL TYPE-2 FUZZY, AND ANN APPROACHES

Kouadria Selman[1], Kouadria Mohamed Abdeldjabbar[2], Berkouk El Madji[3] and Rizoug Nassim[4]

[1]LGEP, Laboratoire de Génie électrique et des plasmas, Université Ibn Khaldoun, Tiaret, Algeria
[2]L2GEGI, Laboratoire de Génie Energétique et Génie Informatique, Université Ibn Khaldoun, Tiaret, Algeria
[3]LCP, Laboratoire de Commande des Processus, Ecole Nationale Polytechnique Algiers, Algeria
[4]ESTACA'Lab, Campus Ouest Parc Universitaire Laval-Changé, France.

[1]https://orcid.org/0000-0001-7179-2146, [2]http://orcid.org/0000-0002-5001-4131
[3]http://orcid.org/0000-0001-9558-2784, [4]http://orcid.org/0000-0002-2151-185X

Email: kouadria.selman@univ-tiaret.dz, mohamedabdeldjabbar.kouadria@univ-mosta.dz, emberkouk@yahoo.fr, nassim.rizoug@estaca.fr

ARTICLE INFO

Article History
Received: July 15, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
wind energy conversion system (WECS),
doubly fed induction generator (DFIG),
Type-1 Fuzzy Logic Control (T1FLC),
Interval Type-2 Fuzzy Logic Control (IT2FLC),
Artificial Neural Networks (ANN).

ABSTRACT

In this study, we investigate and compare three advanced control strategies, Type-1 Fuzzy Logic Control (T1FLC), Interval Type-2 Fuzzy Logic Control (IT2FLC), and Artificial Neural Networks (ANN), applied to Doubly-Fed Induction Generators (DFIGs) in Wind Energy Conversion Systems (WECS). As wind energy plays an increasingly critical role in renewable power generation, efficient and robust control of DFIGs is essential to ensure grid stability, reliability, and optimal energy extraction under variable wind conditions. T1FLC offers simplicity and robustness but may suffer from limited adaptability. IT2FLC improves upon this by managing higher levels of uncertainty and noise through the use of fuzzy sets with an additional degree of freedom. ANN-based control, on the other hand, leverages data-driven learning and non-linear mapping to achieve high performance but may require significant training and tuning. Simulations conducted in MATLAB/Simulink evaluate the performance of each controller under dynamic wind profiles, measuring key metrics such as rotor speed regulation, electromagnetic torque response, and power output stability. This comparative study provides insights for selecting suitable control schemes for DFIG-based wind energy systems under varying operational conditions.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The increasing global focus on sustainable energy solutions has made wind energy a key part of the renewable energy shift. Known for being environmentally friendly, cost-effective, and abundant, wind power plays a major role in decreasing reliance on fossil fuels and reducing greenhouse gas emissions. At the heart of this change are Wind Energy Conversion Systems (WECS), especially those using Doubly-Fed Induction Generators (DFIGs), which provide better control, higher energy efficiency, and flexibility to adapt to changing wind conditions [1], [2]. DFIG-based Wind Energy Conversion Systems (WECS) are widely adopted due to their ability to operate effectively over a wide range of speeds, maximizing power output through advanced control methods like Maximum Power Point Tracking (MPPT). Unlike fixed-speed systems, variable-speed DFIG systems allow for real-time adjustments to rotor speed, boosting energy capture and reducing mechanical stress. However, controlling DFIGs is difficult because of their nonlinear dynamics, the coupling between control variables, and their sensitivity to parameter changes and external disturbances. [3], [4]. To address these challenges, researchers have shifted away from traditional control strategies like Proportional-Integral (PI) controllers. Although these are simple and commonly used, they often lack robustness in uncertain and variable operating conditions. Consequently, there has been a rise in intelligent control methods that can more effectively handle the dynamic and nonlinear characteristics of wind energy systems [5], [6].

Among these, Type-1 Fuzzy Logic Controllers (T1-FLCs) offer improved adaptability by incorporating linguistic rules and expert knowledge to handle system nonlinearities without requiring precise mathematical models. However, their performance can degrade in the presence of significant uncertainties. To address this, Interval Type-2 Fuzzy Logic Controllers (IT2-FLCs) have been developed, extending the fuzzy logic paradigm by modeling uncertainties within membership functions using a Footprint of Uncertainty (FOU). This allows IT2-FLCs to provide enhanced robustness and tracking performance under turbulent wind conditions and parameter variations [7-9]. Parallel to fuzzy logic approaches, Artificial Neural Networks (ANNs) have emerged as a powerful tool for intelligent control. Leveraging their ability to learn complex nonlinear mappings from data, ANNs can adapt to changing dynamics without explicit system modeling, making them highly suitable for unpredictable environments.

However, their effectiveness is closely tied to the quality of training data and the computational resources available [10], [11]. This study presents a comparative analysis of three advanced intelligent control strategies, Type-1 FLC, Interval Type-2 FLC, and ANN applied to DFIG-based wind energy systems. The objective is to evaluate each controller's performance in terms of robustness, tracking accuracy, and computational efficiency under variable wind and system conditions. Through this analysis, the paper aims to identify optimal control methodologies that can enhance the reliability and efficiency of modern wind energy systems, supporting their broader integration into the power grid. Figure 1 illustrates the global scheme of the proposed control strategy, in which three types of controllers are investigated: Type-1 fuzzy logic controller (T1-FLC), Interval type-2 fuzzy logic controller (IT2-FLC), and artificial neural network (ANN) controller.

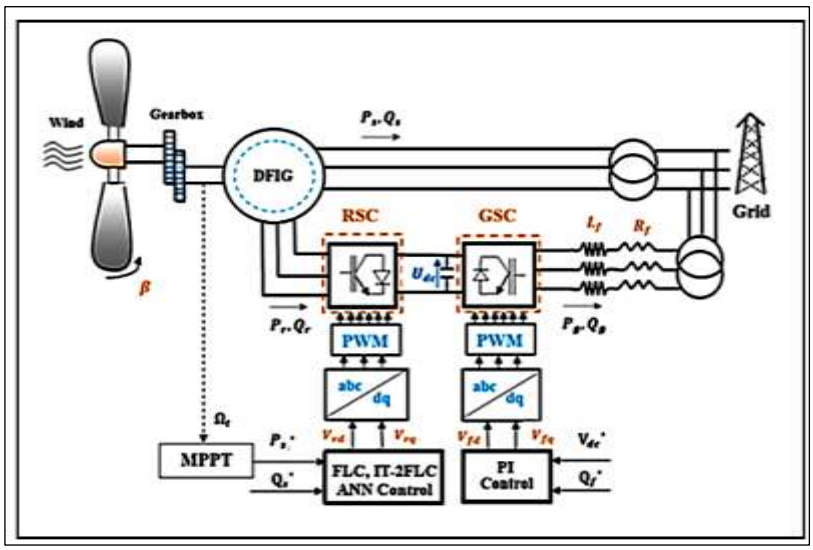


Figure 1: Wind Energy Conversion System configuration.
Source: Authors, (2026).

This paper is structured into six sections. Section 1 provides an overview of wind energy conversion systems and highlights the motivation for applying intelligent control strategies in Doubly-Fed Induction Generator (DFIG)-based systems. Sections 2 and 3 present the modeling of the wind turbine and describe the implementation of a Maximum Power Point Tracking (MPPT) algorithm, respectively. Section 4 focuses on the mathematical modeling of the DFIG. Section 5 presents the simulation results under various operating conditions to evaluate and compare the performance of Type-1 Fuzzy Logic Control (T1-FLC), Interval Type-2 Fuzzy Logic Control (IT2-FLC), and Artificial Neural Network (ANN)-based control strategies. Finally, Section 6 concludes the paper with a summary of the key findings and recommendations for future research.

II. MODELING OF THE WIND GENERATOR

II.1 MODEL Of WIND TURBINE

The wind turbine is the core component of a WECS, and its mechanical power output is determined by its aerodynamic properties as described by the following formula [4],[5],[12].

$$P_w = \frac{1}{2}\rho\pi R^2 v_t^3 C_P(\lambda,\beta) \tag{1}$$

Where:

$$\lambda = \frac{\omega_t . R}{v_t} \tag{2}$$

The aerodynamic torque is commonly represented as follows[13]:

$$T_w = \frac{1}{2}\frac{\rho.\pi.R^3.v_t^2.C_P(\lambda,\beta)}{\lambda} \tag{3}$$

A fundamental equation is used to represent $C_P(\beta,\lambda)$, which is derived from the aerodynamic characteristics of the turbine [13-16]:

$$C_p\left(\lambda,\beta\right)=0.5176.\left(116.\frac{1}{\lambda_i}-0.4.\beta-5\right)e^{-\frac{21}{\lambda_i}}+0.0068.\lambda \tag{4}$$

With

$$\frac{1}{\lambda_i}=\frac{1}{0.08.\beta+\lambda}-\frac{0.035}{1+\beta^3} \tag{5}$$

II.2 MAXIMUM POWER POINT TRACKING TECHNIQUE

MPPT strategies are essential for maximizing power extraction in WECS and improving efficiency. Two main approaches exist: with or without speed control. This discussion focuses on speed control-based MPPT, which regulates turbine rotation to optimize power extraction [4], [5]. Maximum power is achieved when C_P reaches its highest value at a specific λ , corresponding to an optimal rotor speed for each wind speed. However, wind variability can cause deviations from this optimal point, leading to energy losses. An MPPT system continuously optimizes energy transfer to the grid to mitigate these losses [17].

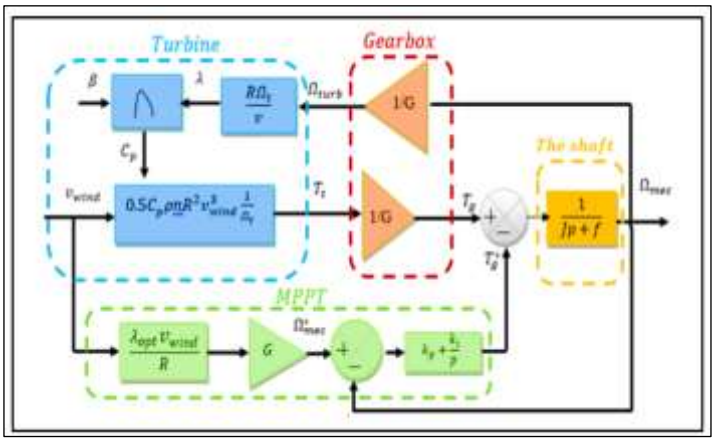


Figure 2: Diagram block of the MPPT extracted with speed control.
Source: Authors, (2026).

The $C_{P\max}=0.48$ is attained when the $\beta=0$ and $\lambda=8.12$ .To sustain this condition, the reference electromagnetic torque is maximized using the given equation [4], [17]:

$$T_{em-ref}=\frac{1}{2}\frac{\rho\pi R_t^5 C_{p\max}}{\lambda_{opt}^3 G^3}\Omega_{mec}^2 \tag{6}$$

Figure 3 presents the $C_P(\beta,\lambda)$ characteristics corresponding to various pitch angle values β [12].

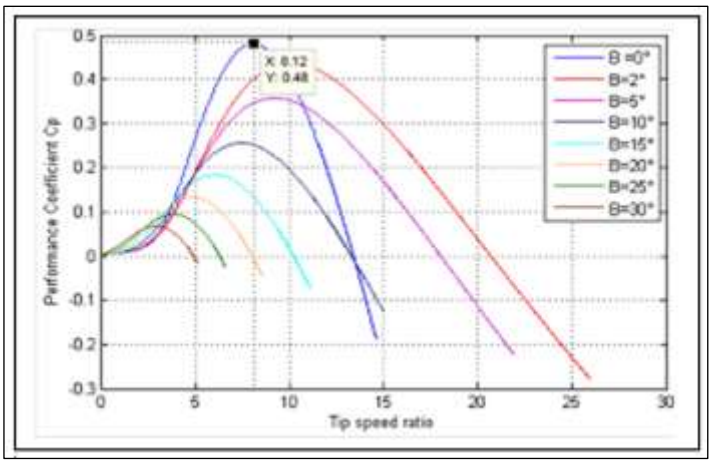


Figure 3: The typical performance coefficient curve.
Source: Authors, (2026).

II.3 DFIG MATHEMATICAL MODELING

The DFIG model is developed using Park's transformation [3-5], [18]. The voltage Equations:

$$\begin{cases} V_{qs} = \dfrac{d\varphi_{qs}}{dt} + R_s I_{qs} - \omega_s \varphi_{ds} \\ V_{ds} = \dfrac{d\varphi_{ds}}{dt} + R_s I_{ds} - \omega_s \varphi_{qs} \\ V_{qr} = \dfrac{d\varphi_{qr}}{dt} + R_r I_{qr} - (\omega_s - \omega_r)\varphi_{dr} \\ V_{dr} = \dfrac{d\varphi_{dr}}{dt} + R_r I_{dr} - (\omega_s - \omega_r)\varphi_{qr} \end{cases} \tag{7}$$

Flux Equations:

$$\begin{cases} \varphi_{ds} = L_m I_{dr} + L_s I_{ds} \\ \varphi_{qs} = L_m I_{qr} + L_s I_{qs} \\ \varphi_{dr} = L_m I_{ds} + L_r I_{dr} \\ \varphi_{qr} = L_m I_{qs} + L_r I_{qr} \end{cases} \tag{8}$$

DFIG electromagnetic torque is:

$$T_{em} = \frac{3}{2} p\left(\varphi_{ds} I_{qs} - \varphi_{qs} I_{ds}\right) \tag{9}$$

P_S and Q_S are determined by:

$$\begin{cases} P_s = \dfrac{3}{2}\left(V_{ds} I_{ds} + V_{qs} I_{qs}\right) \\ Q_s = \dfrac{3}{2}\left(V_{qs} I_{ds} - V_{ds} I_{qs}\right) \end{cases} \tag{10}$$

II.4 DFIG MODEL WITH STATOR FLUX ORIENTATION

The RSC operates in a synchronous reference system with a d-axis aligned to the stator flux, allowing independent control of P and Q power [18]:

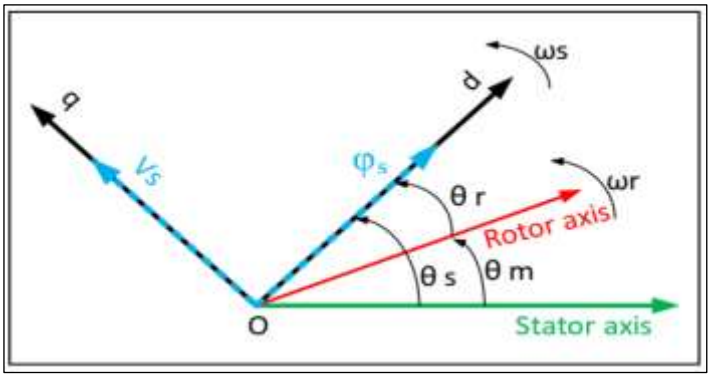


Figure 4: Orientation of the d, q frame.
Source: Authors, (2026).

$$\begin{cases} \varphi_{ds} = \varphi_s \\ \dfrac{d\varphi_{qs}}{dt} = \varphi_{qs} = 0 \end{cases} \tag{11}$$

Apply the following criteria to get the simplified equation for electromagnetic torque:

As a result, we get

$$T_{em} = \frac{3}{2} p \varphi_s I_{qs} \tag{12}$$

Assuming the R_S is ignored, the voltage and flux equations simplify to:

$$\begin{cases} V_{ds} = 0 \\ V_{qs} = \omega_s \varphi_{ds} \end{cases} \tag{13}$$

$$\begin{cases} \varphi_s = L_s I_{ds} + L_m I_{dr} \\ 0 = L_s I_{qs} + L_m I_{qr} \end{cases} \tag{14}$$

Equation (14) defines the correlation between stator and rotor currents.

$$\begin{cases} I_{ds} = \dfrac{\varphi_s}{L_s} - \dfrac{L_m}{L_s} I_{dr} \\ I_{qs} = -\dfrac{L_m}{L_s} I_{qr} \end{cases} \tag{15}$$

P_S and Q_S are given by:

$$\begin{cases} P_s = \dfrac{3}{2} V_s I_{qs} \\ Q_s = \dfrac{3}{2} V_s I_{ds} \end{cases} \tag{16}$$

The electromagnetic torque is determined by integrating equations (12) and (15):

$$T_{em} = -\frac{3}{2} p \frac{L_m}{L_s} \varphi_s I_{qr} \tag{17}$$

By substituting the expressions for the stator currents found in (15), the power equations can be expressed in the following form:

$$\begin{cases} P_s = -\dfrac{3}{2} V_s \dfrac{L_m}{L_s} I_{qr} \\ Q_s = \dfrac{3}{2} V_s \left(\dfrac{\varphi_s}{L_s} - \dfrac{L_m}{L_s} I_{dr} \right) \end{cases} \tag{18}$$

The voltage equations can be represented as follows:

$$\begin{cases} V_{dr} = R_r I_{dr} - g\omega_s \left(L_r - \dfrac{L_m^2}{L_s} \right) I_{qr} \\ V_{qr} = R_r I_{qr} + g\omega_s \left(L_r - \dfrac{L^2_m}{L_s} \right) I_{dr} + g \dfrac{L_m V_s}{L_s} \end{cases} \tag{19}$$

III. INTELLIGENT CONTROL STRATEGIES

III.1 DESCRIPTION OF FUZZY LOGIC SETS

To address the limitations of classical control methods, particularly their reduced performance when actual machine parameters differ from those assumed in the control design, we chose to implement a fuzzy logic controller to manage the behavior of the DFIG wind turbine, focusing primarily on regulating active and reactive power [8].

III.1.1 Overview of Type-1 Fuzzy Logic Sets (T1-Fls)

For any element x within a universe X, a type-1 fuzzy set assigns a membership degree, between 0 and 1, to that element using a function $\mu_A(x)$. We can therefore represent a type-1 fuzzy set as a collection of ordered pairs, where each pair consists of an element and its associated membership degree, as precisely defined by Equation (20). [7-10]:

$$A = \left(x, \mu_A(x) \right) \middle| \forall x \in X \tag{20}$$

Where

$$\mu_A : X \to [0,1]$$

In this definition, $\mu_A(x)$ represents the membership degree of the element $x \in X$ to the set A. In this work, the following notation is used:

$$A(x) = \mu_A(x) \tag{21}$$

For all $x \in X$.

III.1.2 Control of Doubly Fed Induction Generator Using T1-Flc

Figure 5 illustrates the DFIG-based Wind Energy Conversion (WEC) system model simulated using a conventional Type-1 Fuzzy Logic Controller (T1-FLC). Two control loops are presented:

The power control loop regulates the active and reactive power by generating the reference rotor current components I_{rq}^{*} and I_{rd}^{*}, respectively. It employs Type-1 Fuzzy Logic Controllers (T1-FLCs) that process the power errors and their rate of change to provide smooth and adaptive control responses. The current control loop ensures that the actual rotor current components follow the generated references. It also utilizes T1-FLCs to compute the appropriate voltage signals for the rotor-side converter.

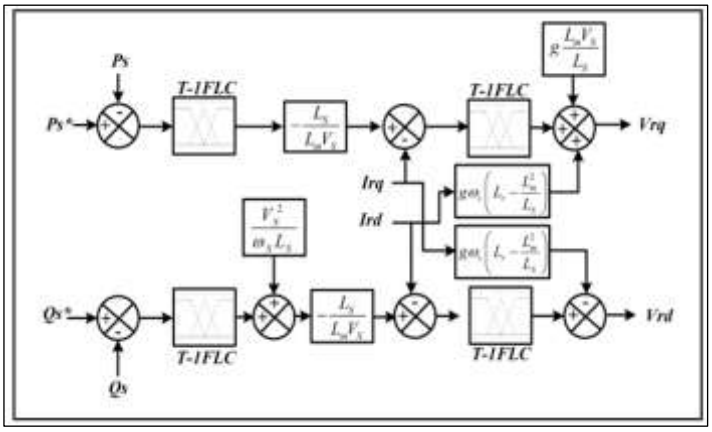


Figure 5: Block diagram illustrating the structure of a type-1 fuzzy logic controller.
Source: Authors, (2026).

III.1.3 Basic Concepts of Interval Type-2 Fuzzy Logic Sets (It-2fls)

According to the definition of Zadeh, A type-2 fuzzy sets are the generalized forms of type-1 fuzzy sets. An interval type-2 fuzzy set, denoted as $\tilde{\tilde{A}}$, is characterized by a type-2 membership function $\mu_{\tilde{\tilde{A}}}(x,u)$ and can be described as [7],[8],[14]-[16],[18]-[22]:

$$\tilde{\tilde{A}} = \left\{ (x,u), \mu_{\tilde{\tilde{A}}}(x,u) \right\} \left| \forall x \in X, \forall u \in J_x^u \subseteq [0,1] \right. \tag{22}$$

In which

$$x \in X, u \in J_x^u \subseteq [0,1] \text{ and } \quad 0 \le \mu_{\tilde{\tilde{A}}}(x,u) \le 1$$

$\left(\tilde{\tilde{A}}\right)$ can be also expressed as:

$$\tilde{\tilde{A}} = \int_{x \in X} \int_{u \in Jx} \mu_{\tilde{\tilde{A}}}(x,u) / (x,u), Jx \subseteq [0,1] \tag{23}$$

Where $\int_{x \in X} \int_{u \in Jx}$ denote union over all admissible input variables x and u.

Here, X is the primary domain of the input variable, x is the value of the input variable, u is the primary grade of a type-2 fuzzy set, and J_x is called the primary membership of a type-2 fuzzy set of x and $\mu_{\tilde{\tilde{A}}}(x,u)$ is the secondary membership function equal to 1.

Another expression for $\left(\tilde{\tilde{A}}\right)$ is:

$$\tilde{\tilde{A}} = \int_{x \in X} \int_{u \in Jx} 1 / (x,u), Jx \subseteq [0,1] \tag{24}$$

$$\tilde{\tilde{A}} : X \to \left\{ [a,b] : 0 \le a \le b \le 1 \right\}$$

The union of all the primary memberships represents the uncertainty about $\left(\tilde{\tilde{A}}\right)$, which is called the footprint of uncertainty (Fou) of $\left(\tilde{\tilde{A}}\right)$ as shown in Figure 3, i.e., [7],[8].

$$Fou\left(\tilde{\tilde{A}}\right) = U_{\forall \in X} J_x = \left\{ (x,u) : u \in J_x \subseteq [0,1] \right\} \tag{25}$$

The (Fou) of $\left(\tilde{\tilde{A}}\right)$ is bounded by two type-1 MFs called upper membership function (UMF) and lower membership function (LMF) as shown in Figure 6. The LMF and UMF are denoted as $\underline{\mu}_{\tilde{\tilde{A}}}(x)$ and $\overline{\mu}_{\tilde{\tilde{A}}}(x)$, respectively, and are defined as follows:

$$\underline{\mu}_{\tilde{A}}(x) = \underline{Fou\left(\tilde{A}\right)}, \forall x \in X \tag{26}$$

$$\overline{\mu}_{\tilde{A}}(x) = \overline{Fou\left(\tilde{A}\right)}, \forall x \in X \tag{27}$$

Note that J_x is an interval set; i.e.,

$$Jx = \left\{(x,u) : u \in \left[\overline{\mu}_{\tilde{A}}(x), \underline{\mu}_{\tilde{A}}(x)\right]\right\} \tag{28}$$

An embedded fuzzy set $\left(\tilde{A}_e\right)$ for a continuous universe of discourse x and u is described as

$$\left(\tilde{A}_e\right) = \int_{x \in X}\left(\frac{[1/u]}{x}\right), u \in J_x \tag{29}$$

The set $\left(\tilde{A}_e\right)$ is embedded in $\left(\tilde{A}\right)$ in such a way that the secondary membership functions (MFs) are always one at each value of x . A large number of such embedded type-1 fuzzy sets (T1-FSs) are combined to form the type-2 fuzzy set (T2-FS), which enables the detailed description of the analytical control surface; the addition of the extra levels of classification gives a much smoother control surface and response. The T2-FS can be considered as a combination of many different type-1 fuzzy sets where each (T1-FS) is embedded to form the (Fou) [7-10],[14],[16].

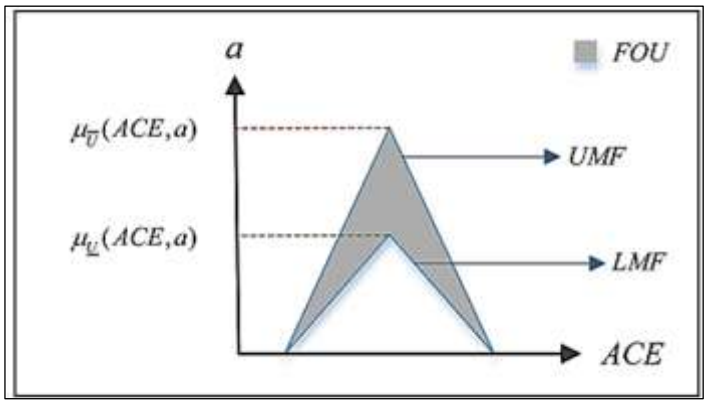


Figure 6: Type-2 fuzzy set with FOU and embedded FS.
Source: [9].

III.1.4 Control of Doubly Fed Induction Generator Using It2-Flc

The Interval type-2 fuzzy logic controller (IT2-FLC) is very useful in circumstances where we need to take into consideration the different uncertainties as the real machine parameters change, and as the controller for a variable speed wind turbine-based on DFIG is used mainly to regulate active and reactive power exchanged with an electrical network. To control a DFIG-based wind energy conversion system easily, we used the independent control of active and reactive power using the stator flux orientation. In the control diagram, the classical rotor currents controllers were replaced by an IT 2-FLC [8],[16].

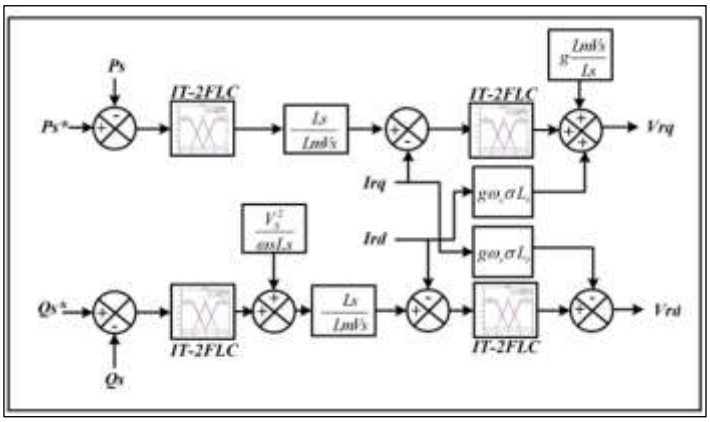


Figure 7: Block diagram illustrating the structure of an Interval type-2 fuzzy logic controller.
Source: Authors, (2026).

The DFIG-based WEC system model was simulated first using a conventional T1-FLC (Figure 5) and then an IT2-FLC. The rule base was the same for both type-1 and type-2 fuzzy logic controllers. The fuzzy control strategy is based on a human operator's experience to interpret a situation and initiate its control action. The scheme of the control of the DFIG-based WEC system using an IT2-FLC is presented in Figure 7.

The core of the type-2 fuzzy logic controller (IT2-FLC) is the type-1 fuzzy logic controller (T1-FLC), as Figure 8 shows. The IT2-FLC is constituted of four functional blocks, the three blocks of type-1 fuzzy and a fourth functional block that ensures the type reducer transforms a type-2 fuzzy set into a type-1 fuzzy set [16]. The structures of IT2-FLS and T1-FLS are shown in Figure 4a and b, respectively [7],[8],[14-16],[19],[20]. The fuzzifier stage is used to translate inputs (real values) to fuzzy values. The inference (reasoning) stage consists of two blocks, the rules base and the inference engine; it works the same way as for type-1 fuzzy systems, except the antecedents' fuzzy sets and the consequent are represented by type-2 fuzzy sets. The process consists of combining the rules base to produce a mapping from input to the output type-2 fuzzy set [7]. It is necessary to calculate the intersection, union and composition of type-2 relations in order to realize this mapping. The type reducer is used to convert all type-2 fuzzy sets into a type-1 fuzzy set on the output. There are several methods to calculate the reduced set, such as joint center, center of sums, height, and center joint, among others [8]. The defuzzification stage translates an output into precise values. A comparison between a T1-FLC and an IT2-FLC is given in the following.

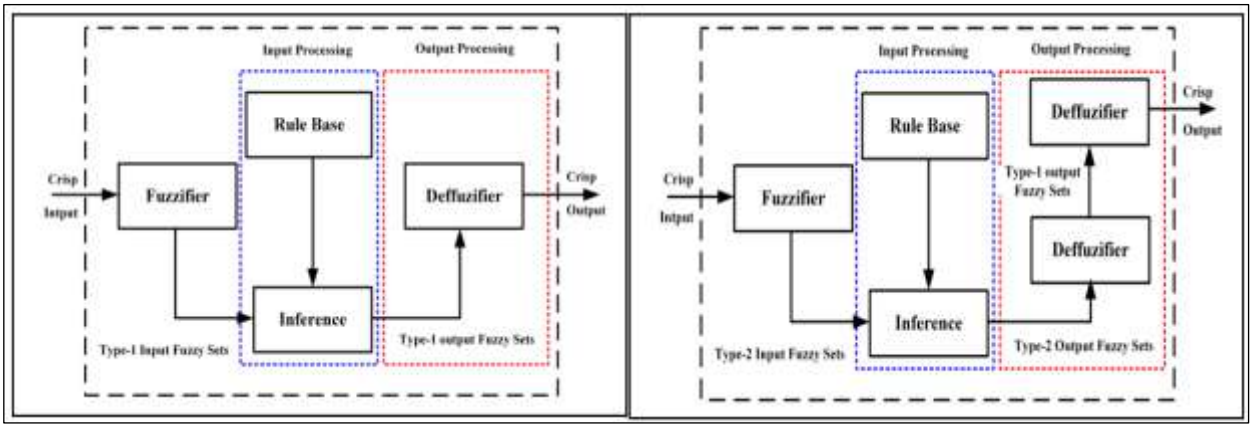


Figure 8: Structures of fuzzy logic control: (a)Type-1 fuzzy logic controller (T1-FLC);
(b) interval type 2 fuzzy logic controller (IT2-FLC).
Source: Authors, (2026).

In this paper, three triangular membership functions were chosen for the IT2-FLC and T1-FLC, because these membership functions are easier to implement in practical hardware, used as the inputs and outputs, which are also shown in Figures 9 and 10, respectively. The maximum and minimum values of the universe of discourse for all inputs and outputs are -1 to +1 [8].]

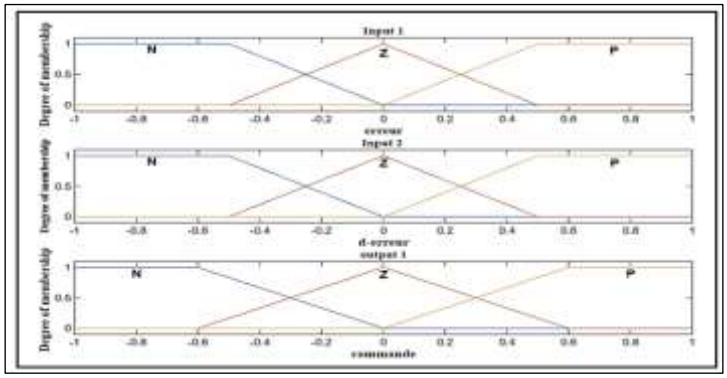


Figure 9: Type-1 fuzzy membership functions (MFs):
MFs corresponding to the input variables 'e' (error)
and 'de' (change in error), as well as the output
variable 'u' of the T1-FLC.
Source: Authors, (2026).

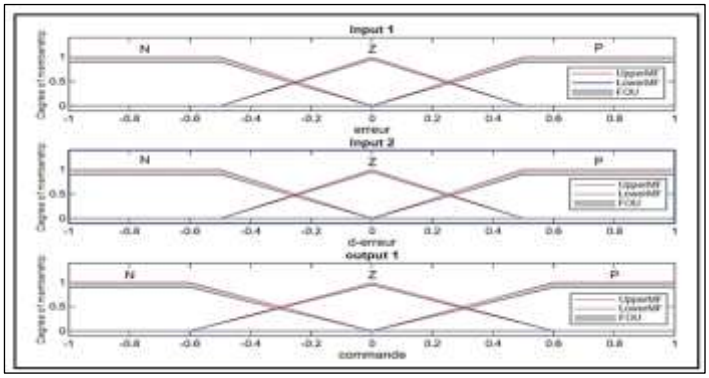


Figure 10: Interval Type-2 fuzzy membership functions (MFs):
MFs corresponding to the input variables 'e' (error)
and 'de' (change in error), as well as the output
variable 'u' of the IT2-FLC.
Source: Authors, (2026).

The controller inputs are error (e) and its derivative (de). The only output of the controller is the control signal (u). Each input and output contain three membership functions, two trapezoidal and one triangular [19]. Input and output membership functions (MFs) of the T1-FLC and IT2-FLC are demonstrated in Figures 5 and 6, respectively. For the optimal performance of the controllers, a 9-rule base was designed for the T1-FLC and IT2-FLC, as shown in Table 1. Referring to this table, N defines the negative. Z is zero while P is positive. The rule base is the core part of the T1-FLC and IT2-FLC design, which is based on process dynamics, experts' knowledge, and experience [23].

Table 1: Fuzzy logic rules.

de / e	N	Z	P
N	N	N	Z
Z	N	Z	P
P	Z	P	P

Source: Authors (2026).

The IT2-FLC was modeled on the MATLAB environment as depicted in Figure 11, and the parameters of the type-2 fuzzy (k1, k2, k3) were determined by the trial-and-error method [7].

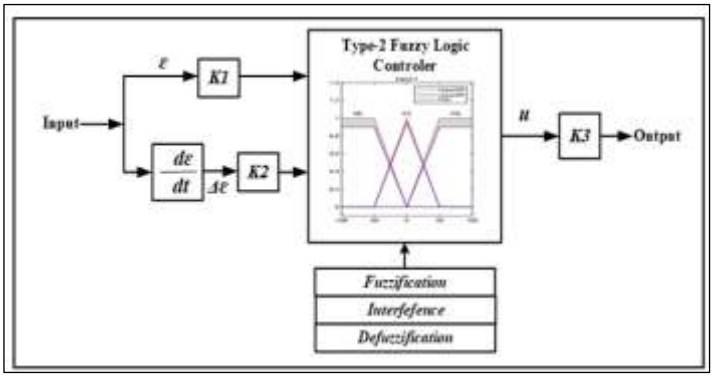


Figure 11: Schematic diagram for the IT2-FLC technique.
Source: Authors, (2026).

III.2. ARTIFICIAL NEURAL NETWORK STRATEGY FOR THE DFIG

An artificial neural network (ANN) is a computational model, inspired by the human brain, made up of interconnected processing units called neurons. These connections have adjustable weights, and each neuron uses an activation function to introduce non-linear behavior. This structure allows the network to learn complex relationships between inputs and outputs. It does this by performing two basic steps: calculating a weighted sum of its inputs and then applying a non-linear activation function to that sum [10],[11].

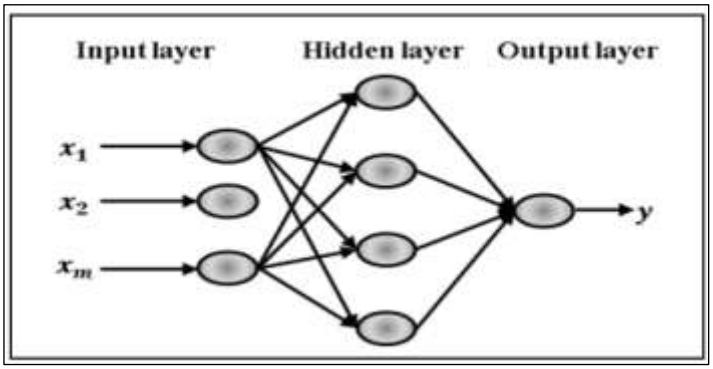


Figure12: Schematic representation of a multilayer Neural Network.
Source [10].

The inherent characteristics of Artificial Neural Networks (ANNs), including parallel processing, learning, and generalization, have established them over the last few decades as one of the most powerful methods for estimation within control systems in various applications. [24], [25]. The following functions represent the operation of the neural network:

$$y = \sum_{k=1}^{n} w_k^2 g\left(\sum_{j=1}^{p} w_{kj}^1 \varphi_j + b_k^1 \right) + b^2 \tag{30}$$

With:

φ_j : ANN input

w_{kj}^1 et b_k^1 : The weights and the bias of the hidden layer

$$g(x) = \frac{2}{1+e^{-2x}} - 1 \tag{31}$$

$g(x)$: The sigmoid function

w_k^2 et b^2 : The weights and the bias of the hidden layer

To take advantage of its advantages, an ANN is proposed in this article to simplify and improve the vector control scheme of the DFIG. The proposed ANN is used to replace the four FLC regulators (T1-FLC and IT2-FLC) in Fig. 4. For the learning of the ANN, it is done offline using a database obtained from the standard vector control diagram. To do this, the inputs and outputs of the FLC controllers are read during the simulation of the vector control for different operating conditions. Then, these data are processed to select a learning base of an appropriate size to ensure a compromise between the sought accuracy and the learning and computing time [26],[27]. Learning is done using the famous retro-propagation algorithm.

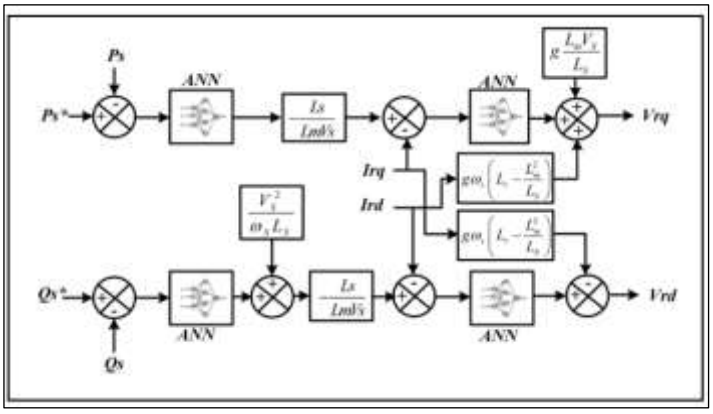


Figure 13: Block diagram illustrating the structure of Interval type-2 fuzzy logic controller.
Source: Authors, (2026).

IV. RESULTS AND DISCUSSIONS

The WECS model utilized in these simulations is based on a 10-kW wind turbine and a 7.5 kW DFIG connected to the grid, and it is implemented in the MATLAB/Simulink environment. The values of the parameters used in the models are listed in the appendix. Figure (14-a) depicts the wind speed profile employed in these simulations. Figures (14-c) and (14-d) show the highest power coefficient $C_{p\max} = 0.48$ and the maximum tip speed ratio $\lambda = 8.12$, for the simulated wind turbine model. The waveforms of optimum angular speed Ω_{opt} and observed angular rotor speed Ω_m of the DFIG are shown in Figure (14-b). The MPPT algorithm determines the reference angular rotor speed Ω_{opt}.

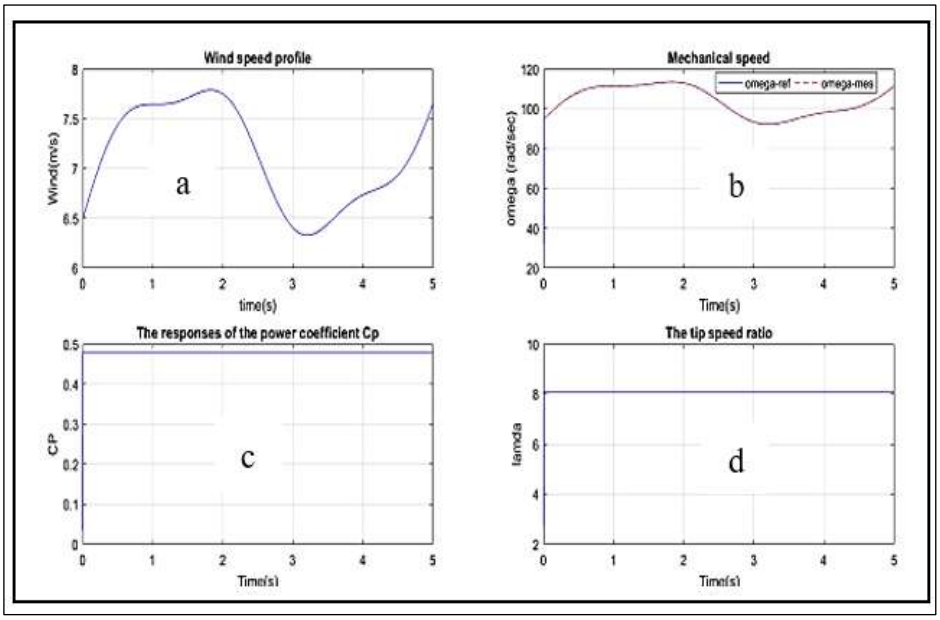


Figure 14: Wind turbine characteristics: (a) Wind speed profile, (b) Mechanical speed, (c) Power Coefficient (Cp), (d) Tip speed ratio (λ).
Source: Authors, (2026).

The simulation results shown in Figures 15, 16, and 17 depict respectively the behavior of the DFIG system under the control of three different intelligent controllers: T1-FLC, IT2-FLC, and ANN. These controllers are used to regulate the active and reactive power generated at the stator, and the control strategy allows for the decoupling of active and reactive power expressions, or equivalently, flux and torque. The simulation confirms that the stator flux follows the reference axis (d) with nearly zero quadrature components, indicating effective decoupling of the machine. The quadrature rotor current (Irq) governs the electromagnetic torque and active power, while the direct rotor current (Ird) manages reactive power exchanged between the stator and the grid. The behavior of these currents directly reflects active and reactive power dynamics, with torque responding rapidly to power demands. Comparatively, T1-FLC demonstrates a moderate transient response with an initial active power spike of approximately (6000 W) and a settling time of (9 ms).

The rotor quadrature current peaks at (20 A) and stabilizes quickly, while the direct current dips to (-20 A) before aligning with its reference of (-15 A) at (0.5 ms). The reactive power shows a dip to (-1000 VAR) and stabilizes to (0 VAR) by (0.7ms), indicating effective voltage support. IT2-FLC, in contrast, exhibits the fastest dynamic performance, with a sharper active power spike (around 16,000 W) but the shortest settling time (around 6 ms). This comes at the cost of a higher (Irq) peak (50 A), showing a more aggressive control effort. The reactive power and (Ird) behaviors are similar to T1-FLC, with stabilization occurring quickly after transients. The ANN controller offers a compromise between these two approaches. It experiences a smaller active power spike (around 8000 W) but with a slower settling time (around 14 ms).

The peak in (Irq) reaches (23.5 A), and (Ird) stabilizes at (-15 A) by (0.4 ms), showing effective control. The reactive power also quickly settles to (0 VAR), confirming proper compensation. In summary, T1-FLC offers balanced and stable regulation with moderate overshoot and fast recovery, making it suitable for smooth operation. IT2-FLC delivers the fastest response and sharpest control action but with higher transient currents and overshoot, which may stress system components. ANN provides robust control with a moderate response and acceptable regulation quality. The choice among these controllers depends on the application's requirements, whether prioritizing speed (IT2-FLC), stability (T1-FLC), or robustness (ANN).]

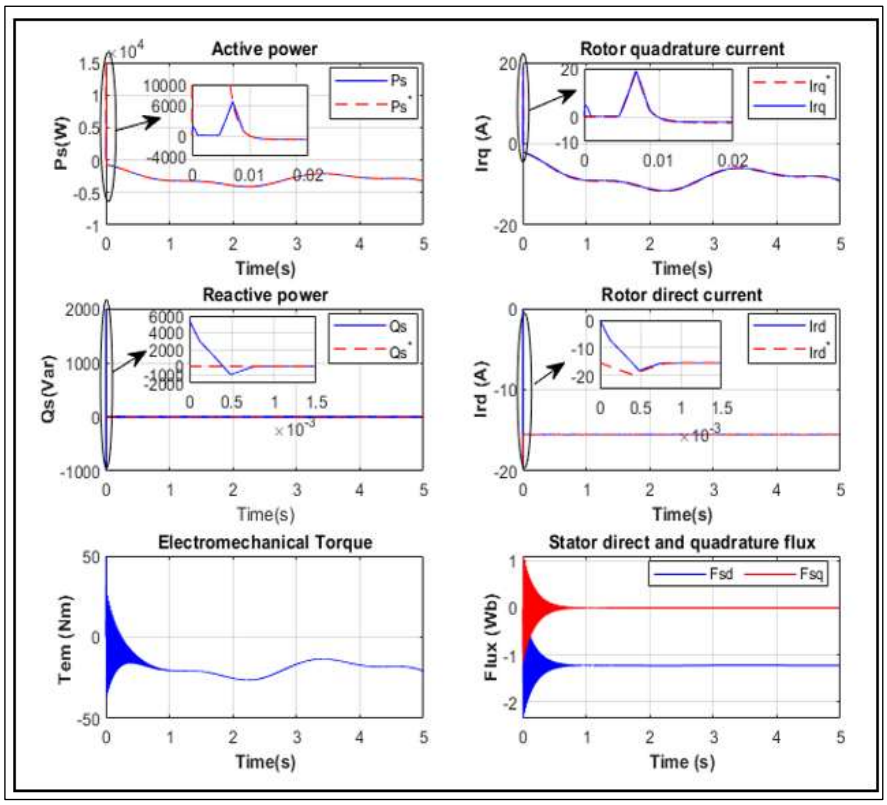


Figure 15: Response of RSC with type 1 fuzzy logic controller.
Source: Authors, (2026).

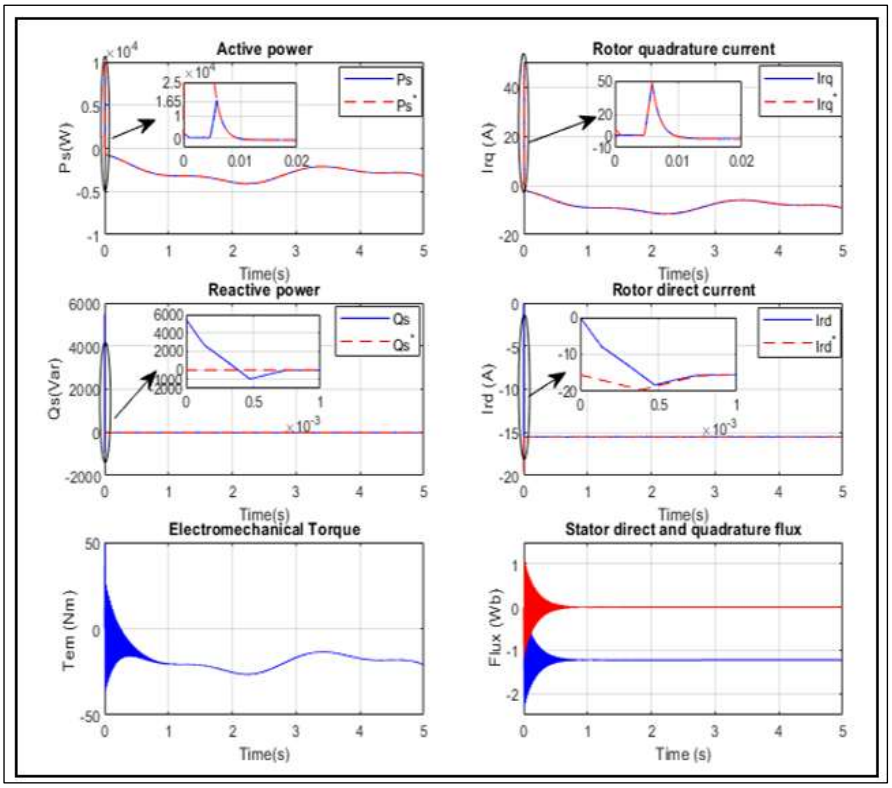


Figure 16: Response of RSC with interval type 2 fuzzy logic controller.
Source: Authors, (2026).

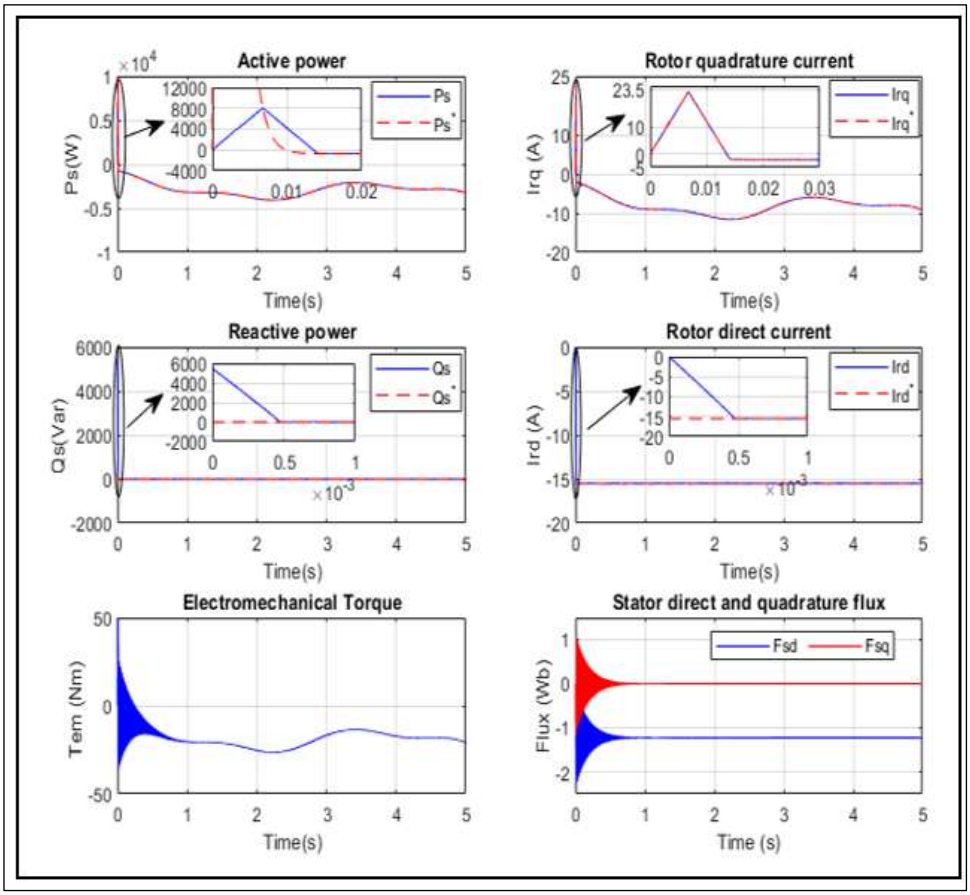


Figure 17: Response of RSC with the ANN controller.
Source: Authors, (2026).

IV.1. QUANTITATIVE COMPARISONS BASED ON ERROR CRITERIA

This test is based on four criteria, which are determined mathematically as follows [6],[21]-[23]. The integral of the squared error (ISE):

$$ISE = \int_0^T e^2(t)\,dt \tag{32}$$

The integral of the absolute value of the error (IAE):

$$IAE = \int_0^T \left|e(t)\right| dt \tag{33}$$

The integral of the time multiplied by the absolute value of the error (ITAE):

$$ITAE = \int_0^T t \cdot \left|e(t)\right| dt \tag{34}$$

The integral of the time multiplied by the value of the squared error (ITSE):

$$ITSE = \int_0^T t \cdot e^2(t)\,dt \tag{35}$$

The criteria are always calculated for a simulation time of Ts = 5 s and power settings equal to those given in the simulation conditions. Table 2 reports a quantitative comparison based on error and time between the proposed control strategies. The results presented in Table 2 show that interval type-2 fuzzy logic control perform the better in terms of minimizing the criteria ISE, IAE, ITAE and ITSE and gives us the lowest values for active and reactive powers and direct and quadrature current, followed respectively by type 1-fuzzy logic control, interval type-2 fuzzy logic control, and artificial neural network.

Table 2: Quantitative comparisons based on error criteria between the developed commands.

G-S	Criterion	Commands Developed for the Considered System		
		T1-FLC	IT2-FLC	ANN
Active powers Ps	ISE	3.811e+8	3.782e+8	3.822e+8
	IAE	1206	1189	1214
	ITAE	62.16	62.14	81.16
	ITSE	5.159e+5	5.084e+5	5.191e+5
Reactive powers Qs	ISE	3328	3229	5104
	IAE	22.24	22.18	30.94
	ITAE	52.77	52.39	74.19
	ITSE	376.1	368.5	668.5
rotor quadrature current Irq	ISE	0.0358	0.0343	0.04318
	IAE	0.06342	0.06335	0.1232
	ITAE	0.148	0.147	0.297
	ITSE	0.002971	0.002917	0.01062
rotor direct current Ird	ISE	0.009932	0.003852	0.01199
	IAE	0.09548	0.08957	0.1087
	ITAE	0.2324	0.2179	0.2679
	ITSE	0.006571	0.005808	0.008872

Source: Authors, (2026).

V. CONCLUSIONS

This study evaluated and contrasted three intelligent control strategies, Type-1 Fuzzy Logic Control (T1-FLC), Interval Type-2 Fuzzy Logic Control (IT2-FLC), and Artificial Neural Networks (ANN), for regulating Doubly Fed Induction Generator (DFIG) based wind turbines. Using a comprehensive system model equipped with MPPT for variable wind speeds, our simulations revealed distinct performance trade-offs. The T1-FLC provided a straightforward control solution with adequate performance under standard operating conditions. In contrast, the ANN controller demonstrated a strong capacity for adaptive learning in response to nonlinear system dynamics; however, its effectiveness is contingent on extensive training data and significant computational power. The IT2-FLC, however, emerged as the superior strategy, offering exceptional robustness against parameter variations and wind turbulence. Its advanced membership structure is uniquely capable of managing and mitigating uncertainty. In summary, IT2-FLC delivered the optimal combination of robustness, precision, and adaptability, establishing it as a highly suitable control method for wind energy applications. Future research will explore hybrid intelligent controllers, specifically the integration of IT2-FLC with advanced nonlinear techniques like second-order sliding mode control, to push performance boundaries further, especially under severe and unpredictable operating scenarios.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Kouadria Selman, Kouadria Mohamed Abdeldjabbar.
Methodology: Kouadria Selman, Kouadria Mohamed Abdeldjabbar.
Investigation: Kouadria Selman, Kouadria Mohamed Abdeldjabbar.
Discussion of results: Kouadria Selman, and Berkouk El Madjid.
Writing – Original Draft: Kouadria Selman.
Writing – Review and Editing: Kouadria Selman, and Berkouk El Madjid.
Resources: Kouadria Mohamed Abdeldjabbar.
Supervision: Berkouk El Madjid and Rizoug Nassim.
Approval of the final text: Kouadria Selman, Kouadria Mohamed Abdeldjabbar and Berkouk El Madjid.

VII. REFERENCES

[1] C. Óh Aiseadha, G. Quinn, R. Connolly, M. Connolly, W. Soon, "Energy and Climate Policy An Evaluation of Global Climate Change Expenditure 2011–2018", Energies 2020, PP 48-39, 2020. https://doi.org/10.3390/en13184839

[2] M. Flores-Granobles, M. Saeys, "Minimizing CO2 emissions with renewable energy: A comparative study of emerging technologies in the steel industry", Energy & Environmental Science, vol. 13, no 7, p. 1923-1932., 2020. https://doi.org/10.1039/D0EE00787K

[3] S. Ouhssain, H Chojaa, Y. Aljarhizi, E. A. Ibrahmi, A Maarif, M. A. Mossa "Enhancing the Performance of a Wind Turbine Based DFIG Generation System Using an Effective ANFIS Control Technique" International Journal of Robotics and Control Systems Vol. 4, no. 4, 2024, pp. 1617-1640 https://doi.org/10.31763/ijrcs.v4i4.1451

[4] S. Kouadria, Y.Messlem, E. M. Berkouk, "Sliding mode control of the active and reactive power of DFIG for variable-speed wind energy conversion system". In: 3rd International Renewable and Sustainable Energy Conference (IRSEC), 2015, p. 1–8.

[5] S. Kouadria, E. M. Berkouk, Y Messlem and M. Denaï. "Improved control strategy of DFIG-based wind turbines using direct torque and direct power control techniques." Journal of Renewable and Sustainable Energy Vol 10, p 043306, 2018. https://doi.org/10.1063/1.5023739

[6] R. Rouabhi, A. Herizi and A. Djerioui, "Performance of Robust Type-2 Fuzzy Sliding Mode Control Compared to Various Conventional Controls of Doubly-Fed Induction Generator for Wind Power Conversion Systems" Energies, vol 17 no 15, p 3778, 2024. https://doi.org/10.3390/en17153778

[7] A. Milles, E. Merabet, H. Benbouhenni, N. Debdouche, I. Colak "Robust control technique for wind turbine system with interval type-2 fuzzy strategy on a dual star induction generator" Energy Reports Vol 11, 2715–2736, 2024, https://doi.org/10.1016/j.egyr.2024.01.060.

[8] A. V. Hemeyine, A. Abbou, A. Bakouri, M. Mokhlis and S. M. o. M. E. Moustapha "A Robust Interval Type-2 Fuzzy Logic Controller for Variable Speed Wind Turbines Based on a Doubly Fed Induction Generator" Inventions, vol 6, no 21, pp 1-14, 2021, https://doi.org/10.3390/inventions6020021.

[9] S. Shirali, S. Z. Moghaddam, M. Aliasghary "An interval Type-2 fuzzy Fractional-Order PD-PI controller for frequency stabilization of islanded microgrids optimized with CO algorithm" Electrical Power and Energy Systems vol 164, p110422, 2025, https://doi.org/10.1016/j.ijepes.2024.110422.

[10] B. Boudjellal, T. Benslimane "Active and Reactive Powers Control of DFIG Based WECS Using PI Controller and Artificial Neural Network Based Controller" International Information and Engineering Technology Association Vol. 93, no. 1, pp. 31-38, 2020, https://doi.org/10.18280/mmc_a.931-405

[11] M. Hallouz, N. Kabache, S. Moulahoum‡, S. Kouadria "Neural Network Based Field Oriented Control for Doubly-Fed Induction Generator" International Journal of Smart Grid Vol.2, No.3, 2018, https://doi.org/10.20508/ijsmartgrid.v2i3.18.g18

[12] S. Kouadria,; S. Belfedhal,; Y. Messlem,; E. M. Berkouk, "Study and control of wind energy conversion system (WECS) based on the doubly fed induction generator (DFIG) connected to the grid. In: Ninth International Conference on Ecological Vehicles and Renewable Energies (EVER), 2014, p. 1–7.

[13] Saad M. Alwash, Osama Qasim Jumah Al-Thahab , Shamam F. Alwash Modeling and Control Strategies for DFIG in Wind Turbines: A Comparative Analysis of SPWM, THIPWM, and SVPWM Techniques International Information and Engineering Technology Association Vol. 56, no. 6, pp. 963-972, 2023, https://doi.org/10.18280/jesa.560607.

[14] I. Allali, B. Dehiba "A Comparative Study of Interval Type-2 and Type-1 Fuzzy Sliding Mode in Controlling DFIG-Based Wind Energy Conversion System Nigerian Journal of Technological Development, vol. 22, no.3, , pp 101 -110, 2025, *doi:*http://dx.doi.org/10.63746/njtd.v22i3.2818.

[15] K. A. Naika, C. P. Guptab, E. Fernandezb "Design and implementation of interval type-2 fuzzy logic-PI based adaptive controller for DFIG based wind energy system" Electrical Power and Energy Systems vol 115 p 105468, 2020, https://doi.org/10.1016/j.ijepes.2019.105468.

[16] S. G. Bruno, D. E. A. Mansour, A. Nada, T. F. Megahed, "Hybrid ANN-Based MPPT Control for DFIG Wind Systems Using Type-2 Fuzzy Logic and Super-Twisting Sliding Mode Control" Smart Grids and Sustainable Energy, vol. 10, no 2, p. 32, 2025, https://doi.org/10.1007/s40866-025-00261-5.

[17] B. Kaima and A. Boukhelifa, "Output Power Control of a Wind Energy Conversion System Based on a Doubly Fed Induction Generator," International Renewable and Sustainable Energy Conference (IRSEC),.pp. 292– 297, 2013.

[18] E. O. Zouggar, S. Chaouch, D. O. Abdeslam, A. L Abdelhamid" Sliding Control with Fuzzy Type-2 Controller of Wind Energy System Based on Doubly Fed Induction Generator" International Information and Engineering Technology Association Vol. 18, no. 2, pp. 137-146, 2019, https://doi.org/ 10.18280/i2m.180207.

[19]. M. Latha Maheswari, D. Nagarajan, J. Kavikumar, S. Broumi, "Triangular interval type-2 fuzzy soft set and its application". Complex Intell. Syst., vol 6, p 531–544, 2020, https://doi.org/10.1007/s40747-020-00151-6.

[20]. S.K. Raju, G.N. Pillai, "Design and Implementation of Type-2 Fuzzy Logic Controller for DFIG-Based Wind Energy Systems in Distribution Networks". IEEE Trans. Sustain. Energy, Vol 7, p 345–353, 2015., https://doi.org/10.1109/TSTE.2015.2496170.

[21] S. K. Raju, G.N. Pillai "Design and real time implementation of type-2 fuzzy vector control for DFIG based wind generators" Renewable Energy, vol. 88, p. 40-50, 2016, https://doi.org/10.1016/j.renene.2015.11.006.

[22]. M.F. Hamza, H. J. Yap, I.A. Choudhury, "Recent advances on the use of meta-heuristic optimization algorithms to optimize the type-2 fuzzy logic systems in intelligent control". Neural Computing and Applications, vol. 28, no 5, p. 979-999, 2017, https://doi.org/10.1007/s00521-015-2111-9.

[23] R. Rouabhi, A. Zemmit, A. Herizi, O. Moussa, S. Djeriou, "Hybrid type-2 fuzzy backstepping control of doubly fed induction generator for wind energy conversion systems", Journal of the Brazilian Society of Mechanical Sciences and Engineering, vol. 47, no 1, p. 24, 2025, https://doi.org/10.1007/s40430-024-05293-z.

[24] Y. Djeriri, A. Meroufel, et M. Allam, "Artificial neural network-based robust tracking control for doubly fed induction generator used in wind energy conversion systems", J. Adv. Res. Sci. Technol., vol. 2, no 1, p. 173-181, 2015.

[25] B. Azeem et al., "Robust neural network scheme for generator side converter of doubly fed induction generator", in 2017 International Symposium on Recent Advances in Electrical Engineering (RAEE), p. 1-6, 2017.

[26] N. Kabache, S. Moulahoum, K. Sebaa, et H. Houassine, "Neural network-based input output feedback control of induction motor", in Optimization of Electrical and Electronic Equipment (OPTIM), 2012 13th International Conference on, Romania, pp. 578–583, 2012.

[27] Y. Djeriri, "Direct power control based artificial neural networks of doubly fed induction generator for wind energy conversion system application", J. Adv. Res. Sci. Technol., vol. 5, no 1, p. 592–603, 2018.

VIII. APPENDIX

Table 3:Wind Turbine parameters.

Nominal power (P)	10KW
Diameter (D)	3m
Number of blades	3
Gearbox (G)	5.4
Inertial (J)	0.042 kg.m^2
Viscous coefficient (f)	0.017 N.m. s^{-1}

Source: Authors (2026).

Table 4: Doubly Fed Induction Generator parameters.

Nominal power (P)	7.5KW
Nominal voltage (V)	230/380 V
Nominal speed (N)	1500 rpm
Stator resistance (Rs)	0.455 Ω
Rotor resistance (Rr)	0.62 Ω
Stator inductance (Ls)	0.084 H
Rotor inductance (Lr)	0.081 H
Main inductance (Lm)	0.078 H
Pairs of pole number (p)	2
Inertial (J)	0.3125 $Kg.m^2$
Viscous coefficient (f)	$6.73.10^{-3}$ $N.m.s^{-1}$

Source: Authors (2026).

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 16-23. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2646

ISSN ONLINE: 2447-

RESEARCH ARTICLE OPEN ACCESS

MULTI-OBJECTIVE OPTIMIZATION OF CMOS FOLDED-CASCODE OTAS USING DIVERSE EVOLUTIONARY ALGORITHMS

Pankaj P. Prajapati[1], Alpesh M. Patel[2*], Anilkumar J. Kshatriya[3], Kirit V. Patel[4], Ghanshyamkumar Sah[5], Chintan Dave[6], Kinjal R. Sheth[7], Devendra H. Patel[8]

[1,6,7]Electronics and Communication, Vishwakarma Government. Engineering College, Ahmedabad, Gujarat, India.
[2,3,4]Electronics and Communication, Government Engineering College, Gandhinagar, Gujarat, India
[5]Electronics and Communication, L. D. College of Engineering, Ahmedabad, Gujarat, India.
[8]Electronics and Communication, Government Engineering College, Patan, Gujarat, India.

[1]http://orcid.org/0000-0001-6731-4264, [2]http://orcid.org/0000-0001-8846-4771, [3]http://orcid.org/0000-0002-9044-0442,
[4]http://orcid.org/ 000-0002-1280-9937, [5]http://orcid.org/0009-0009-4049-2410, [6]http://orcid.org/0009-0002-1082-7928,
[7]http://orcid.org/0000-0003-0132-3533, [8]http://orcid.org/ 0009-0000-8222-6684

Email: pankajprajapati@vgecg.ac.in, *am_patel@gtu.edu.in, anil@gecg28.ac.in, kirit@gecg28.ac.in, ghanshyam@ldce.ac.in, cb_dave@gtu.edu.in, kinjalsheth@vgecg.ac.in, pateldevendrah@gmail.com

ARTICLE INFO

Article History
Received: September 2, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
EA,
DE algorithm,
PSO algorithm,
CS algorithm,
Optimization.

ABSTRACT

The manual development of CMOS-based analog circuits continues to pose significant challenges, largely due to device scaling, process-induced variations, and the nonlinear interactions among multiple interdependent design parameters. These complexities necessitate the advancement of automated methodologies to improve efficiency, scalability, and reliability in analog design. In this study, the Cuckoo Search (CS) algorithm is investigated as a robust evolutionary framework for the automated sizing of CMOS analog circuits, with particular emphasis on the Folded Cascode Operational Transconductance Amplifier (FOTA) implemented in both 0.18 µm and 0.35 µm CMOS technologies. To provide a rigorous comparative analysis, the performance of CS is benchmarked against two widely employed evolutionary algorithms, namely Differential Evolution (DE) and Particle Swarm Optimization (PSO). All algorithms were implemented in the C programming language, interfaced with the NGSPICE simulator, and executed on an Intel® Core™ i5 processor (2.40 GHz, 8 GB RAM) under the Ubuntu operating system. The experimental findings reveal that CS consistently exhibits superior performance in terms of success rate, convergence reliability, and robustness. Specifically, in the case of the 0.35 µm FOTA design, CS achieved convergence to all target specifications in 9 out of 10 independent runs, outperforming DE (eight successes) and PSO (five successes). Comparable superiority of CS was also observed for the 0.18 µm process, thereby validating its potential across multiple technology nodes. This study demonstrates that CS not only provides superior optimization outcomes compared to DE and PSO but also establishes itself as a highly effective and reliable evolutionary framework for automated analog circuit design. The cross-technology validation and performance benchmarking highlight its practical relevance in advancing analog design automation

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

CMOS analog circuits form the foundation of systems ranging from precision sensors to high-speed communication links, yet their design remains demanding. Device-level variations, parasitic effects, and nonlinear behavior make analog optimization far more intricate than its digital counterpart. At the same time, designers must reconcile requirements for gain, bandwidth, stability, efficiency, and area, a combination that calls for rigorous multi-objective optimization methods.Among circuit-level challenges, transistor sizing is

the most critical. The folded-cascode operational transconductance amplifier (OTA) exemplifies these difficulties: its architecture resolves voltage headroom constraints while retaining high gain, offering wide input range and strong power-supply rejection. Yet its performance is highly sensitive to transistor dimensions, bias allocation, and parasitic interactions. Trade-offs among gain, bandwidth, slew rate, and power dissipation are further complicated by process variation.This study explores evolutionary algorithms (EAs) as a framework for transistor-level optimization, emphasizing their integration with high-fidelity circuit models. Through systematic evaluation, it demonstrates the capacity of EAs to handle the complexity of analog design spaces and to deliver balanced solutions across conflicting performance metrics, thereby advancing the automation and reliability of CMOS analog circuit design.

II. LITERATURE REVIEW

PSO has been widely used for automated sizing of CMOS analog circuits in SPICE-in-the-loop frameworks due to its simplicity, exploratory power, and reliable convergence in complex design spaces. Research has mainly focused on addressing its stagnation tendency and sensitivity to parameter settings. Shreeharsha et al. [1] developed an enhanced-velocity PSO (EV-PSO) with adaptive inertia weight and acceleration factors for CMOS op-amp sizing. The method encouraged broad exploration in the early phase and shifted to precise refinement later, leading to quicker convergence and a larger set of practical solutions than the basic PSO. In high-frequency circuit design, where parasitics and layout effects make optimization difficult, Liu et al. [2] applied a randomized PSO (R-PSO) to the co-design of a 60 GHz subharmonic mixer. Constraint handling is equally critical, since many analog design requirements—such as minimum phase margin or maximum slew rate—are imposed as strict limits rather than soft targets.

Bouali et al. [3] developed a constraint-repaired PSO variant that incorporates adaptive neighborhooding to dynamically adjust the influence topology of particles. According to studies on particle swarm optimization (PSO), which also examined the influence of its hyperparameters, the swarm can vary or converge too early due to improper values of inertia or learning coefficients [4]. These issues have led to the introduction of parallel PSO implementations that employ distributed evaluation of possible solutions. Particle swarm optimization (PSO) has been adapted to many-objective frameworks as the number of these objectives increases. According to Shu et al. [5], PSO works well in problems with a small number of objectives, but as dimensionality rises, its efficacy decreases.In order to overcome this constraint, hybrid approaches that integrate PSO with Pareto-ranking schemes or decomposition-based techniques have been put forth, enabling the algorithm to maintain both population diversity and convergence.

Differential Evolution (DE) has also shown reliable performance in CMOS optimization, especially when the design space is highly irregular, multimodal, and tightly constrained [6]. Zhao et al. [7] demonstrated DE's capability by combining it with a deterministic direct search for fully automated op-amp synthesis. In multi-objective contexts, advanced DE variants such as GDE3 have been employed in conjunction with surrogate models to reduce computational cost. Vişan et al. [8] demonstrated a Gaussian process-assisted GDE3 for optimizing a folded-cascode OTA, achieving wide Pareto front coverage with a significantly reduced number of SPICE simulations. Recently, DE has also been embedded within Bayesian Optimization frameworks to address budget-constrained optimization problems [9]. Cuckoo Search (CS), inspired by the brood parasitism behavior of certain cuckoo species and characterized by Lévy flight-based random walks, has emerged as a powerful alternative for analog sizing problems that demand both aggressive exploration and the ability to fine-tune within narrow feasible corridors.

The heavy-tailed step distribution inherent to Lévy flights enables CS to traverse wide regions of the design space, thereby avoiding premature convergence a common challenge in constrained analog design tasks. In ultra-low-voltage design, where transistor sizing is tightly constrained by near-threshold operation, Fortes et al. [10] applied a standard CS algorithm to the design of a 0.25 V bulk-driven Miller OTA in 130 nm CMOS. CS has also been adapted for parasitic-aware layout optimization, as demonstrated by Abouelyazid et al. [11]. Their modified CS incorporated placement-and-routing-aware performance estimation, enabling the optimizer to maintain post-layout performance metrics closely aligned with pre-layout targets. A broader methodological assessment by Abualigah et al. [12] underscored CS's suitability for highly nonconvex analog design problems, emphasizing its ability to balance exploration and exploitation dynamically without extensive parameter tuning.

This low-sensitivity to parameter configuration is particularly valuable in industrial CAD flows, where rapid algorithm deployment is often prioritized over prolonged parameter calibration. Z. Xu et al. [13] showed that deterministic multi-objective methods can equal or even outperform evolutionary approaches on smaller analog design tasks. Lberni et al. [14] employed deep neural network surrogates to model the nonlinear input–output mapping of analog circuits, embedding these models within the evolutionary loop. This approach reduced the reliance on computationally expensive SPICE evaluations while preserving high-fidelity optimization, enabling larger design spaces to be explored under fixed computational budgets. Complementary enhancements such as yield-aware optimization [15] and feasibility-biased initial population generation [16] have been increasingly incorporated into hybrid EA frameworks.

II.1 MAJOR GAPS IN THE LITERATURE:

- Evolutionary algorithms are effective for analog circuit sizing but struggle with many-objective problems, strict design constraints, and process variability.
- A scalable and unified optimization framework that addresses these challenges is still lacking.
- Current methods rely heavily on SPICE-in-the-loop evaluations and lack smooth integration into industrial CAD flows.
- There is a need for a holistic approach that unites evolutionary search, surrogate modeling, and deterministic refinement to enable practical, automation-ready analog circuit design.

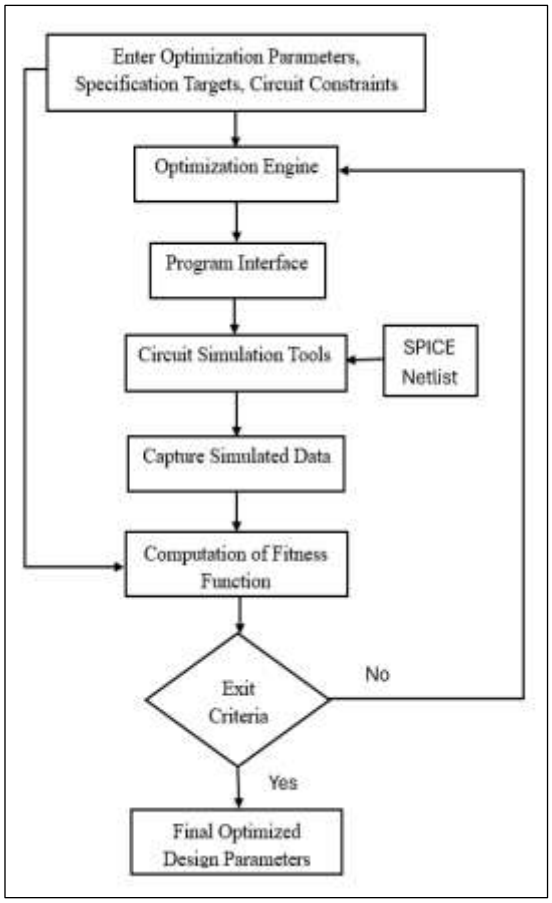


Figure 1: Adopted framework for automated circuit design methodology.
Source: Authors, (2026).

III. MATERIALS AND METHODS

The schematic representation of the automated CMOS analog circuit design framework, as depicted in Figure 1, constitutes a subset of a broader design automation environment. In this framework, the circuit topology is selected beforehand and treated as fixed. The system combines an evolutionary optimization engine with a SPICE-based simulator to synthesize CMOS analog circuits that meet defined performance targets. The optimizer systematically tunes transistor dimensions and bias conditions, while each candidate solution is validated through Ngspice simulations [17]. At every iteration, a parameterized SPICE netlist is generated to capture the electrical behavior of the design. MOS devices are constrained to operate in saturation, verified through bias point analysis and voltage transfer characteristics. The netlist is passed to the simulator to ensure device-level consistency, and a dedicated interface coordinates data flow between the optimizer and SPICE. Optimization begins with EA initialization, defining design variables, bounds, and objectives, followed by population generation and evaluation. This iterative co-simulation process enables detailed exploration of the design space and guides convergence toward high-fidelity solutions that satisfy analog performance requirements. Based on the outcomes produced by the circuit simulator, the fitness function of the circuit is quantitatively evaluated according to the following mathematical formulation [18].

$$Fitness\ Function\ = \sqrt{\sum_{j=1}^{D}\left(\frac{Specification_{desired} - Specification_{simulated}}{Specification_{desired}}\right)_{j}^{2}} \tag{1}$$

Here, D denotes the total number of targeted design specifications. This formulation essentially computes the Root Mean Square (RMS) error, thereby attributing equal weighting to all specifications, irrespective of their individual magnitudes or sensitivities. Consequently, the optimizer is compelled to satisfy each specification with comparable emphasis, ensuring a balanced trade-off among potentially conflicting performance objectives. Following the evaluation of the fitness function, the process proceeds to the termination check. The optimization cycle is concluded only if the pre-defined stopping conditions are fulfilled. In this work, the termination criteria are established as either (i) achieving a fitness function value below a stringent threshold of 1×10^{-6}, or (ii) reaching the upper bound of 100 iterations. Should neither condition be met, the evolutionary algorithm (EA) generates a new candidate set of design parameters, and the cycle is repeated.

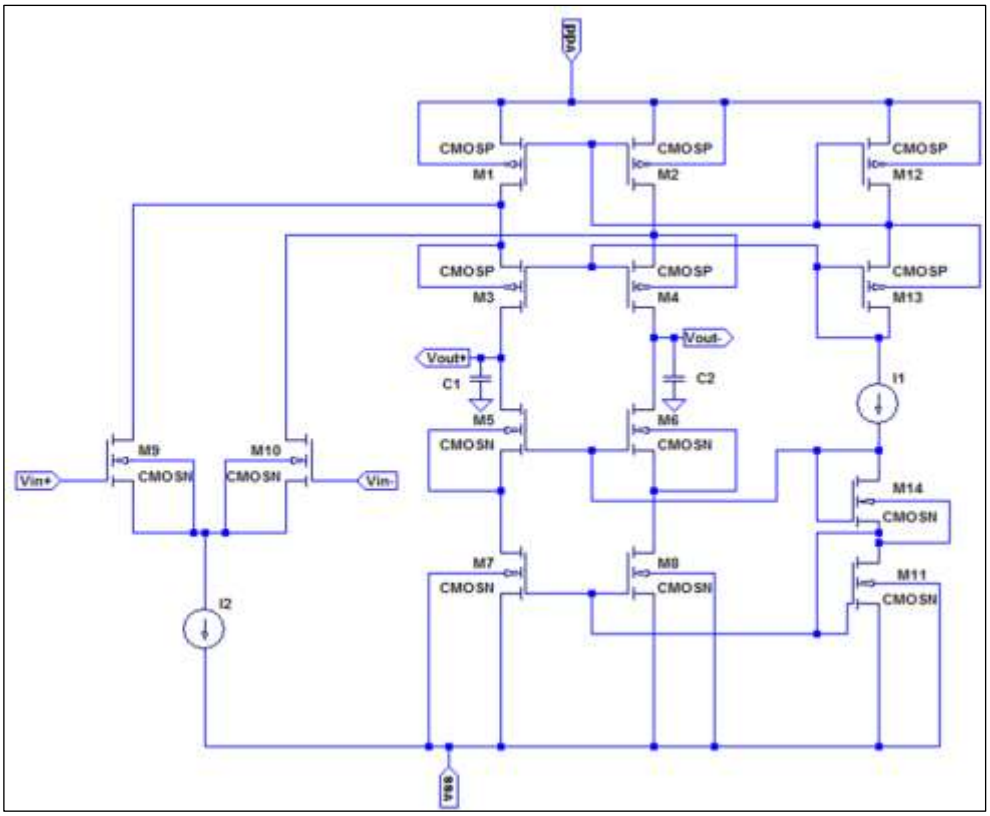


Figure 2: Folded Cascode OTA.
Source: Authors, (2026).

This iterative refinement embodies the optimizer's primary objective: to progressively reduce the fitness function value at each successive iteration until convergence to the desired solution is realized. The optimizer's performance depends on factors such as hyperparameter tuning, design space complexity, variable ranges, and the number of specifications to be met. This work applies three metaheuristic algorithms PSO, DE, and CS to optimize CMOS analog circuits, enabling a comparative assessment of their efficiency, robustness, and suitability for design automation.

IV. RESULTS AND DISCUSSIONS

The circuit diagram of CMOS Folded Cascode OTA (FOTA) is shown in Figure 2 [19]. More theoretical and mathematical analysis of this circuit can be found in [19], [20]. We have considered the width of M1, M2, and M12 transistors as W1 = W2 = W12, the width of M3 and M4 transistors as W3 = W4, the width of M5 and M6 transistors as W5 = W6, the width of M7 and M8 transistors as W7 = W8, the width of M9 and M10 transistors as W9 = W10, the width of M11 transistor as W11, the width of M13 transistor as W13, and the width of M14 transistors as W14 for this circuit. We have set the length of all transistors as L = 1 µm. So, this circuit has nine design parameters namely W1 = W2 = W12, W3 = W4, W5 = W6, W7 = W8, W9 = W10, W11, W13, W14, and I_{bias}. The circuit is optimized to drive the load capacitor of 0.1 pF. The supply voltage is set to ±1.8 V and ±2 V for both 0.18 µm and 0.35 µm CMOS technologies respectively.

Table 1: Search space and optimized parameters by different EAs for FOTA using 0.35 µm CMOS technology.

Sr. No.	Design Parameters	Search Space of Design Parameters	Optimized values of Design Parameters				
			PSO	DE	CS	[20]	[19]
1	$W_{1,2,12}$ (µm)	1 µm to 200 µm	196.24	200	121.31	25.6	28.5
2	$W_{3,4}$ (µm)	1 µm to 200 µm	94.6	93.77	35.11	12.8	49.9
3	$W_{5,6}$ (µm)	1 µm to 200 µm	163.95	200	143.37	4.5	16.4
4	$W_{7,8}$ (µm)	1 µm to 200 µm	163.95	200	143.37	4.5	34.2
5	$W_{9,10}$ (µm)	1 µm to 200 µm	195.34	200	164.99	49.2	49.85
6	W_{11} (µm)	1 µm to 200 µm	200	200	71.37	9	68.4
7	W_{13} (µm)	1 µm to 200 µm	200	200	100.21	25.6	99.8
8	W_{14} (µm)	1 µm to 200 µm	187.98	200	75.07	9	32.8
9	I_{bias} (µA)	1 µA to 200 µA	116.08	133.42	130.54	27.5	30

Source: Authors, (2026).

Table 2 presents a comparative summary of both the targeted specifications and the achieved simulated values corresponding to the optimized design parameters obtained through the application of different evolutionary algorithms (EAs) to this circuit. For this design case, a total of twelve distinct performance specifications are imposed, namely: open-loop voltage gain (Av), unity-gain bandwidth (UGB), phase margin (PM), positive power supply rejection ratio (+PSSR), negative power supply rejection ratio (–PSSR), rail-to-rail swing ratio (RSR), full swing ratio (FSR), common-mode rejection ratio (CMRR), power dissipation (P_{diss}), input-referred noise spectral density (Nin) at 10 kHz, integrated input-referred noise (N_{int}) across the frequency band from 1 Hz to 100 kHz, and the Total MOS Transistor Area (TTA).

The search space and optimized values achieved by different EAs for the design parameters of this circuit using 0.35 µm CMOS technology are found better than [19] and [20] listed in Table 1. Previous investigations have also explored the optimization of this circuit topology under comparable CMOS technologies. Vural et al. [21] reported the application of the standard Particle Swarm Optimization (PSO) algorithm for the design of this circuit, while Mallick et al. [22] extended this line of research by employing a Craziness-based PSO (CRPSO) variant to further enhance the exploration capabilities of the swarm and mitigate premature convergence issues. In comparison, the present results demonstrate improved performance over those reported by Vural et al. [21] using standard PSO and by Mallick et al. [22] with the CS variant, as summarized in Table 2.

Table 2: Desired specifications and obtained specifications by different EAs for FOTA using 0.35 µm CMOS technology.

Sr. No.	Specifications	Desired value	Obtained Specifications				
			PSO	DE	CS	[21]	[22]
1	A_V (dB)	> 70	74.15	73.53	74.45	76	85. 5
2	UGB (MHz)	> 100	157.70	156.39	219.42	420	537
3	PM (°)	> 45	83.64	64.97	81.79	--	---
4	+ve PSSR (dB)	> 65	68.82	68.93	70.03	57.68	79
5	- ve PSSR (dB)	> 75	84.85	85.16	83.24	57.68	79
6	RSR (V/µs)	> 40	155.89	90.40	85.49	200	290
7	FSR (V/µs)	> 40	53.83	54.33	67.85	200	290
8	CMRR (dB)	> 80	117.50	129.88	110.97	59.23	61.75
9	P_{diss} (mw)	< 1	0.739	0.851	0.880	0.66	0.696
10	N_{in} (nV^2/Hz)	< 1e-6	7.75 e-8	6.75 e-8	5.23 e-8	--	---
11	N_{op} (nV^2/Hz)	< 1e-3	4.06 e-4	3.44 e-4	4.26 e-4	--	---
12	TTA (μm^2)	< 3000	2412.38	2587.54	1584.26	262.4	1155.37

Source: Authors, (2026).

Table 3: Performance of different EAs for optimization of FOTA using 0.35 µm CMOS technology.

Algorithm	$SD_{fitness}$	$Iter_{avg}$	FE_{avg}	S_{rate}	T_{sim} (s)
DE	**0.030888**	60	1857	8	3835
PSO	0.233654	56.40	**1692**	5	**3393**
CS	0.090025	56.20	3402	**9**	6442

Source: Authors, (2026).

The desired design specifications, along with the corresponding values obtained through parameter optimization using different evolutionary algorithms (EAs) for this circuit in a 0.35 µm CMOS technology, are summarized in Table 1. The comparative performance of three well-established optimization techniques PSO, DE and CS was systematically evaluated through 10 independent experimental runs, as presented in Table 3. The CS algorithm showed the greatest robustness, meeting all target specifications in 9 of 10 runs, compared with eight for DE and five for PSO. These results position CS as a promising tool for complex analog design tasks involving high-dimensional parameter spaces and tight performance constraints. For the 0.18 µm CMOS implementation, Table 4 lists the parameter search ranges alongside the optimized values obtained by each algorithm. The results highlight the efficiency with which different algorithms navigate the high-dimensional search landscape to satisfy stringent performance specifications.

The desired design specifications are presented in Table 5 for the circuit implemented in 0.18 µm CMOS technology. A detailed performance comparison of PSO, DE, and CS across 10 independent optimization trials is provided in Table 6. The results reveal that both CS and DE consistently achieved convergence to all targeted specifications in every run. PSO achieved full specification compliance in only 4 of 10 runs, reflecting weaker robustness in navigating the circuit's design space. Evaluation of CS showed the most efficient convergence, requiring fewer iterations on average than both DE and PSO. This efficiency reflects CS's stronger balance between exploration and exploitation in high-dimensional spaces. While DE demanded slightly more computation than CS, it remained more reliable and efficient than PSO.

Table **Erro! Nenhum texto com o estilo especificado foi encontrado no documento.**: Search space and optimized parameters by different EAs for FOTA using 0.18 µm CMOS technology.

Sr. No.	Design Parameters	Search Space of Design Parameters	Optimized values of Design Parameters		
			PSO	DE	CS
1	$W_{1,2,12}$ (µm)	1 µm to 200 µm	110.47	200	44.81
2	$W_{3,4}$ (µm)	1 µm to 200 µm	99.66	74.85	3.79
3	$W_{5,6}$ (µm)	1 µm to 200 µm	172.18	200	174.84
4	$W_{7,8}$ (µm)	1 µm to 200 µm	172.18	200	174.84
5	$W_{9,10}$ (µm)	1 µm to 200 µm	108.2	200	63.72
6	W_{11} (µm)	1 µm to 200 µm	200	149.89	7.44
7	W_{13} (µm)	1 µm to 200 µm	159.5	200	181.08
8	W_{14} (µm)	1 µm to 200 µm	175.58	200	96.17
11	I_{bias} (µA)	1 µA to 200 µA	142.84	130.32	149.55

Source: Authors, (2026).

Table 5: Desired specifications and obtained specifications by different EAs for FOTA using 0.18 µm CMOS technology.

Sr. No.	Specifications	Desired value	Obtained Specifications		
			PSO	DE	CS
1	A_V (dB)	> 70	77.71	77.86	79.03
2	UGB (MHz)	> 100	167.19	146.54	194.02
3	PM (°)	> 45	76.72	70.13	80.93
4	+ve PSSR (dB)	> 65	77.92	77.09	78.47
5	- ve PSSR (dB)	> 75	77.35	78.55	79.67
6	RSR (V/µs)	> 10	88.09	166.63	40.43
7	FSR (V/µs)	> 10	52.95	41.01	57.69
8	CMRR (dB)	> 80	137.17	139.92	140.35
9	P_{diss} (mw)	< 1	0.867	0.779	0.843
10	N_{in} (nV^2/Hz)	1e-6	4.06 e-8	5.11 e-8	2.51 e-8
11	N_{op} (nV^2/Hz)	1e-3	2.79 e-4	2.96 e-4	2.77 e-4
12	TTA (μm^2)	< 3000	1970.93	2499.59	1253.5

Source: Authors, (2026).

Table 6: Performance of different EAs for the optimization of FOTA using 0.18 µm CMOS technology.

Algorithm	$SD_{fitness}$	$Iter_{avg}$	FE_{avg}	S_{rate}	T_{sim} (s)
DE	0.0	38	1170	10	2428
PSO	0.218294	69	2067	4	4088
CS	0.0	38.30	2328	10	4493

Source: Authors, (2026).

Taken together, these findings establish that while both CS and DE demonstrate strong potential as optimization frameworks for analog circuit design in deep-submicron CMOS technologies, CS holds a distinctive advantage due to its consistent reliability coupled with faster convergence. The weaker performance of PSO in this study highlights the shortcomings of swarm-based heuristics when applied to highly constrained, nonlinear, and multi-objective design problems. By comparison, the results point to CS and DE as more effective optimization strategies for analog design automation at the 0.18 µm technology node.

V. CONCLUSIONS

This work examined the use of the Cuckoo Search (CS) algorithm for automated design of CMOS analog circuits, focusing on the Folded Cascode OTA in 0.18 µm and 0.35 µm technologies. When compared to Particle Swarm Optimization (PSO) and Differential Evolution (DE), CS consistently produced higher success rates, faster convergence, and more robustness under demanding design constraints. In 9 of 10 trials, CS converged for the 0.35 µm process, surpassing DE with 8 and PSO with 5. The effectiveness of CS across technology scales is confirmed by similar results at 0.18 µm.The study emphasizes DE's relative competitiveness while highlighting PSO's shortcomings in complex analogue design. However, CS was able to strike an improved balance between exploration and improvement, which made it a dependable method for circuit automation in the face of complexity and scaling issues. Because of its proven effectiveness, it can be integrated into next-generation CAD platforms for mixed-signal SoCs, which can shorten development cycles and increase design productivity. Future directions include extension to advanced nodes, multi-objective strategies that explicitly recognize trade-offs in power, area, and noise, and hybrid approaches that integrate CS with other heuristics.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Pankaj P. Prajapati, Alpesh M. Patel, Kirit V. Patel, and Ghanshyamkumar Sah.
Methodology: Pankaj P. Prajapati, Anilkumar J. Kshatriya, Kirit V. Patel, and Kinjal R. Sheth.
Investigation: Pankaj P. Prajapati, Alpesh M. Patel, Chintan Dave, and Devendra H. Patel.
Discussion of results: Anilkumar J. Kshatriya, Alpesh M. Patel, Kirit V. Patel, and Devendra H. Patel.
Writing – Original Draft: Pankaj P. Prajapati, Alpesh M. Patel, Ghanshyamkumar Sah, and Anilkumar J. Kshatriya.
Writing – Review and Editing: Alpesh M. Patel, Anilkumar J. Kshatriya, Kinjal R. Sheth and Ghanshyamkumar Sah.
Resources: Pankaj P. Prajapati, Kirit V. Patel, Chintan Dave, Devendra H. Patel and Kinjal R. Sheth.
Supervision: Pankaj P. Prajapati, Ghanshyamkumar Sah, Chintan Dave and Devendra H. Patel.
Approval of the final text: Alpesh M. Patel, Pankaj P. Prajapati, Chintan Dave, and Kinjal R. Sheth.

VII. REFERENCES

[1] K. G. Shreeharsha, R. K. Siddharth, M. H. Vasantha and Y. B. N. Kumar, "Partition Bound Random Number-Based Particle Swarm Optimization for Analog Circuit Sizing," in IEEE Access, vol. 11, pp. 123577-123588, 2023, doi: 10.1109/ACCESS.2023.3329698

[2] W. Liu, Z. Wang, N. Zeng, Y. Yuan, F. E. Alsaadi, and X. Liu., "A novel randomised particle swarm optimizer," International Journal of Machine Learning and Cybernetics, vol. 12, no. 2, pp. 529–540, 2021, doi: 10.1007/s13042-020-01186-4.

[3] H. Bouali, B. Benhala, and M. Guerbaoui, "Multi-objective optimization of CMOS low noise amplifier through nature-inspired swarm intelligence," Bulletin of Electrical Engineering and Informatics, vol. 12, no. 5, pp. 2824–2836, Oct. 2023, doi: 10.11591/eei.v12i5.5512.

[4] H. Xu, Q. Deng, Z. Zhang, and S. Lin, "A hybrid differential evolution particle swarm optimization algorithm based on dynamic strategies," Scientific Reports, vol. 15, Art. no. 4518, Feb. 2025, doi: 10.1038/s41598-024-82648-5.

[5] X. Shu, Y. Liu, and Q. Zhang, "A novel multiobjective particle swarm optimization combining hypercube and distance," Scientific Programming, vol. 2022, Art. ID 9448419, 2022, doi: 10.1155/2022/9448419.

[6] D. Joshi, S. Dash, S. Reddy, R. Manigilla, and G. Trivedi, "Multi-objective hybrid particle swarm optimization and its application to analog and RF circuit optimization," Circuits, Systems, and Signal Processing, vol. 42, no. 8, pp. 4443–4469, 2023, doi: 10.1007/s00034-023-02342-1.

[7] Z. Zhao, J. Liu, and L. Zhang, "Acceleration for efficient automated generation of operational amplifiers," Sensors (Basel), vol. 24, no. 12, art. no. 3990, 20 Jun. 2024, doi: 10.3390/s24123990.

[8] C. Vişan, O. Pascu, M. Stănescu, E.-D. Şandru, C. Diaconu, A. Buzo, G. Pelz, and H. Cucu, "Automated circuit sizing with multi-objective optimization based on differential evolution and Bayesian inference," Knowledge-Based Systems, vol. 258, p. 109987, 2022, doi: 10.1016/j.knosys.2022.109987.

[9] L. Kouhalvandi, O. Ceylan, and S. Ozoguz, "Optimization techniques for analog and RF circuit designs: An overview," Analog Integrated Circuits and Signal Processing, vol. 106, no. 3, pp. 511–524, 2021, doi: 10.1007/s10470-020-01733-7.

[10] A. Fortes, F. A. Quirino, L. A. da Silva, and A. Girardi, "Low power bulk-driven OTA design optimization using cuckoo search algorithm," Analog Integrated Circuits and Signal Processing, vol. 106, no. 1, pp. 99–109, 2020, doi: 10.1007/s10470-020-01632-x.

[11] M. S. Abouelyazid, S. Hammouda, and Y. Ismail, "RC parasitic-aware layout analysis and routing optimization methodology," IEEE Access, vol. 10, pp. 92740–92759, 2022, doi: 10.1109/ACCESS.2022.3203077.

[12] L. Abualigah, A. Ababneh, A. M. Ikotun, R. Abu Zitar, A. R. Alsoud, N. Khodadadi, A. E. Ezugwu, E. S. Hanandeh, and H. Jia, "A survey of cuckoo search algorithm: optimizer and new applications," in Metaheuristic Optimization Algorithms, Elsevier, 2024, pp. 45–57, doi: 10.1016/B978-0-443-13925-3.00018-2.

[13] Z. Xu, Z. Zhao, and J. Liu, "Deterministic multi-objective optimization of analog circuits," Electronics, vol. 13, no. 13, p. 2510, 2024, doi: 10.3390/electronics13132510.

[14] A. Lberni, M. A. Marktani, A. Ahaitouf, and A. Ahaitouf, "Analog circuit sizing based on evolutionary algorithms and deep learning," Expert Syst. Appl., vol. 237, art. no. 121480, Mar. 2024, doi: 10.1016/j.eswa.2023.121480.

[15] A. Canelas, F. Passos, N. Lourenço, R. Martins, E. Roca, R. Castro-Lopez, N. Horta, and F. V. Fernández, "Hierarchical yield-aware synthesis methodology covering device-, circuit-, and system-level for radiofrequency ICs," IEEE Access, vol. 9, pp. 124152–124164, 2021, doi: 10.1109/ACCESS.2021.3110758.

[16] M. Zuo and Y. Xue, "Population feasibility state guided autonomous constrained multi-objective evolutionary optimization," Mathematics, vol. 12, no. 6, art. no. 913, Mar. 2024, doi: 10.3390/math12060913.

[17] P. Nenzi and H. Vogt, Ngspice User's Manual. 2014. [Online]. Available: http://ngspice.sourceforge.net/

[18] Fakhfakh, M., Cooren, Y., Sallem, A., Loulou, M., & Siarry, P., "Analog Circuit Design Optimization through the Particle Swarm Optimization Technique," Analog Integrated Circuits and Signal Processing, vol. 63, no. 1, pp. 71–82, Jan. 2010. doi: 10.1007/s10470-009-9361-3.

[19] A. Kaur and M. Kaur, "A comprehensive survey of test functions for evaluating the performance of particle swarm optimization algorithm," Int. J. Hybrid Inf. Technol., vol. 8, no. 5, pp. 97–104, 2015. doi: 10.14257/ijhit.2015.8.5.10.

[20] M. M. Ali and A. Törn, "Population set-based global optimization algorithms: Some modifications and numerical studies," Comput. Oper. Res., vol. 31, no. 10, pp. 1703–1725, 2004. doi: 10.1016/S0305-0548(03)00116-3.

[21] R. A. Vural and T. Yildirim, "Analog circuit sizing via swarm intelligence," Int. J. Electron. Commun., vol. 66, no. 9, pp. 732–740, 2012

[22] S. Mallick, R. Kar, S. P. Ghoshal, and D. Mandal, “Optimal sizing and design of CMOS analogue amplifier circuits using craziness-based particle swarm optimization,” Int. J. Numer. Model. Electron. networks, devices fields, vol. 29, no. 5, pp. 943–966, 2016. doi: 10.1002/jnm.2155. ;

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 24-34. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2679

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

COMPARATIVE STUDY OF HMM AND BPNN IN DETECTING CORONA DISCHARGE ON 20 KV CUBICLE BASED ON VOLTAGE AND SOUND

Christiono Christiono*[1], Miftahul Fikri[2] and Syamsir Abduh[3]

[1, 2, 3]Institur Teknologi PLN. Jakarta, Indonesia.

[1]https://orcid.org/0000-0001-5457-9976, [2]https://orcid.org/0000-0002-9454-9197, [3]https://orcid.org/0009-0000-7984-8511

Email:*christiono@itpln.ac.id

ARTICLE INFO

Article History
Received: September 6, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
corona discharge,
20kV cubicle,
Voice
Hidden Markov Model (HMM),
Back Propagation Neural Network (BPNN),

ABSTRACT

Corona discharge is a common disturbance in 20 kV cubicles and may lead to insulation degradation, power losses, and potential equipment failure. Early detection is therefore essential to support preventive maintenance and enhance distribution system reliability. This study presents a comparative analysis of the Hidden Markov Model (HMM) and Backpropagation Neural Network (BPNN) for identifying corona discharge based on acoustic signals and voltage variations. Acoustic data were recorded using a needle–rod configuration and processed through Linear Predictive Coding (LPC) to obtain cepstral features. The classification results show that HMM provides high accuracy, achieving 100% in both training and testing for noise-based clustering and 84.44% in voltage-based testing. Meanwhile, BPNN demonstrates stable performance with training and testing accuracies of 95.93% and 82.35% for voltage-based clustering, and 95.69% and 100% for noise-based clustering, respectively. Overall, the findings indicate that although HMM performs more consistently for noise-based classification, BPNN provides competitive accuracy and better adaptability for early detection of corona discharge in 20 kV cubicles. These results support the development of intelligent monitoring systems for improving insulation condition assessment and reducing potential failures in medium-voltage distribution networks.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Greater Jakarta is the center of electricity supply. PLN Distribution Jaya and UP3Cikokol have more nearly 7 million consumers, 11,564 MVA of installed power, and 12,426 distribution substations, according to PT PLN (Persero) figures from 2006. Cubicles and feeders are needed for three times as many distribution substations as there are substations.[1]**.** So, if done manually, the inspection in the context of maintenance will be very difficult. A voltage of 20 kV is applied to these cubicles, which can cause corona discharge symptoms and, if left unchecked, cause electrical failure or short circuits, especially under rain or humid environments[2]**.** If the repair is done immediately, the problem can be solved, but you can only find out the condition of the cubicle by visiting it[3]. Many people consider that corona discharge in a medium voltage cubicle (TM) is not enough to cause corona discharge.

However, the fact is that corona discharge can cause a flash jump, causing the cubicle to be unable to deliver electrical power [4]. Due to the high humidity level in Indonesia which allows air ionization to occur easily, the corona discharge problem in the TM cubicle is a unique and uncommon problem because the magnitude of the voltage (20 kV) is technically very difficult to cause corona discharge [5]. For example, PT. PLN (Persero) Jaya and Tangerang sustain more than 300 units of TM cube damage annually. Most of this damage is caused by the release of corona. Therefore, there needs to be a monitoring system that can see data from the entire room but can be accessed from the control center. With the latest condition data, technicians can immediately carry out maintenance to prevent more severe corona discharge, which can cause flashovers[6]. Figure 1 shows the structure and condition after cubicle damage.



(a) (b)

Figure 1: Cubicle Construction 20 kV Voltage (a) visible inside cubicle (b) visible part of the entire cubicle.
Source: Authors, (2026).

Figure 1 shows a 20 kV cubicle that suffered an insulation failure due to a previously unknown corona discharge phenomenon that caused permanent damage to its insulator. Part (a) shows the arrangement of the insulators inside the cubicle, and part (b) shows the overall cubic units; Both have suffered permanent damage due to insulation failure.

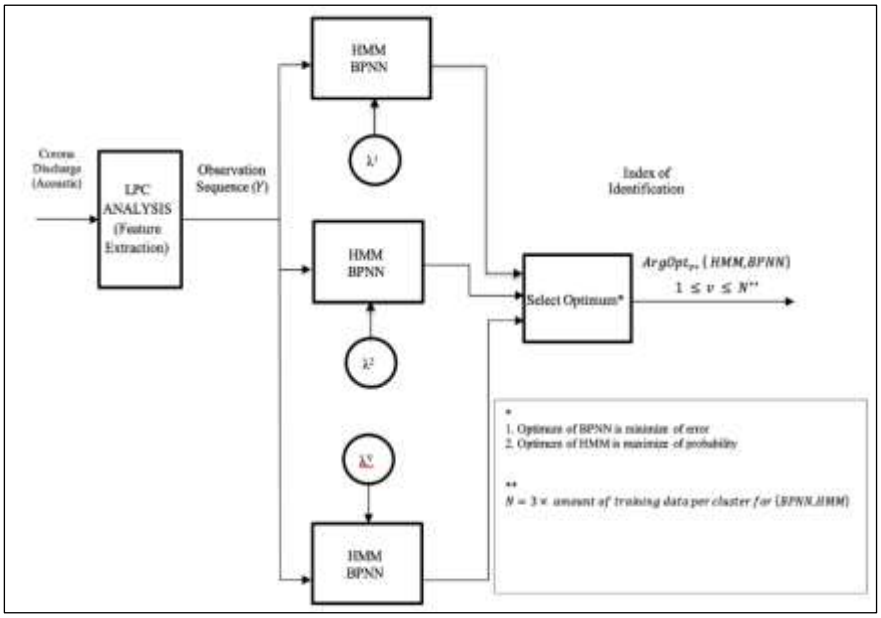


Figure 2: Acoustic Identification Flow Diagram of Corona Discharge Based on Voltage and Noise.
Source: Authors, (2026).

Any kind of medium-voltage or high-voltage electrode may cause corona discharge symptoms in specific situations.[7]. This is because an ionization process that can generate electrons or positive ions at 20 kV cubic insulation occurs when the electrode is subjected to an electric field. At some point, this collection of electrons or positive ions will form into an avalanche electron, which can then produce sound. The more the point group, the more interference the electrode produces. [8]. The noise level can be measured as a parameter of the magnitude of the level of corona discharge symptoms. This can be done by raising the voltage from 20 kV to 33 kV (less than the breakdown voltage).

Some corona discharge studies such as [9]. studying the loose corona on the transmission wire using loose partial measurement,[10]. Talking about corona discharge using unmanned aerial system technology,[11]. talking about how to classify corona discharge based on the pattern of partial discharge, [12]. detects corona discharge sound using fast Fourier transform. Then, using wavelet transformation to identify corona discharge signals based on the level of contamination or pollution. This signal is compared to a partial discharge detector measuring device. Recently, corona discharge sound clustering using geometric distance (ED) has been carried out.[13][14], using fuzzy c-mean (FcM) [15][16].and compare ED, FcM and hidden Markov model (HMM) methods [17].

However, AI uses backpropagation neural network (BPNN) to identify corona discharge sounds. In this study, clustering of the corona discharge phenomenon based on sound was carried out as a first step. This clustering is intended to identify early isolation failures in 20 kV cubicles using BPNN. The characteristic extraction of the recorded corona discharge sound is computed using the linear predictive coding (LPC) approach, which has long been recognized and proven to be dependable for obtaining characteristic extraction from any sound.After the extraction of these traits is obtained, matching is carried out using the BPNN and HMM methods to obtain clustering that is expected to be more accurate.

II. METHODOLOGY

In the acoustic observation of corona discharge, this study refers to IEEE Standard for High-Voltage Testing Techniques 4 2013-1995 IEEE [17] and IEC TS 62478:2016 High voltage test techniques - Measurement of partial discharges by electromagnetic and acoustic using PD data collection method. On the other hand, the use of needle-rod electrodes in vertical testing and microphone placement is at a distance of 3cm. With a standard frequency ranging between 40 and 200 kHz, this study used only 40 kHz. Each classification was done ten times in order to detect the acoustic phenomenon of corona ejection based on voltage and hissing sound. Preliminary observations show that the lowest penetration voltage is 34.3 kV, so the base voltage is clustered into 3 including: low cluster 20–24 kV, medium cluster 25–29 kV, and high cluster 30–33 kV, while sizzle noise is classified into pure Corona Discharger, Corona Discharger with hissing noise, and pure hissing noise. The humidity ranges from 70% to 95%, while the indoor temperature ranges from 27.5°C to 35.3°C. Data was gathered at the PLN Institute of Technology Jakarta's High Voltage Laboratory over a one-year observation period (February 2021–February 2022). The MATLAB Program version 2024b is then used to process the collected data. Figure 2 shows a flow diagram of the clustering of corona discharge phenomena based on acoustics.

II.1 PREDICTIVE LINEAR CODE

According to him, the LPC approach is a long-standing speech recognition methodology. [17]. The LPC method's benefits include its ease of use and compatibility with a wide range of devices. The two primary parts of the LPC method are encoding and decoding. While decoding modifies the recorded sound, encoding breaks down sound signals into their component elements. Here, both procedures are explained. Figure 3 [17].

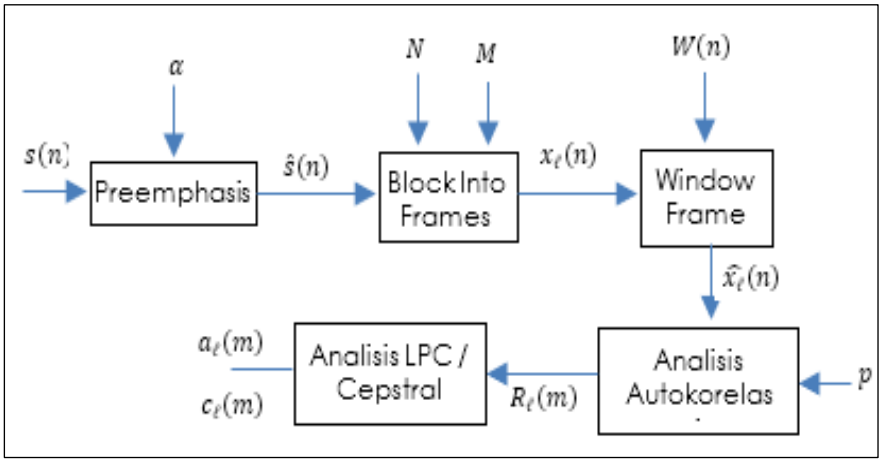


Figure 3: LPC Process as Acoustic Corona Release Feature Extraction.
Soucer: [13], [14], [16], [17].

with

$$\tilde{s}(n) = s(n) - \tilde{a}\, s(n-1), \tag{1}$$

where the acoustic data/signal,$sn$is the amount of data in an acoustic signal

$$0.9 \leq a \leq 1.$$
$$x_\ell(n) = \tilde{s}(M\ell + n) \tag{2}$$

where shows the number of frames,$M\ell = 1,2, \dots, M$.

$$\tilde{x}_\ell(n) = x_\ell(n)\, w(n), \tag{3}$$

Where$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), 0 \leq n \leq N-1$

$$r_\ell(m) = \sum_{n=0}^{N-1-m} \tilde{x}_\ell(n)\, \tilde{x}_\ell(n+m) \tag{4}$$

Where$m = 0,1, \dots, p$.

$$a_m = \alpha_m^{(p)} \tag{5}$$

with$1 \leq m \leq p$

$$E^{(0)} = r(0)$$

$$k_i = \frac{\left\{r(i) - \sum_{j=1}^{L-1} \alpha_j^{(i-1)} r(|i-j|)\right\}}{E^{(i-1)}}$$

$$\alpha_i^{(i)} = k_i$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}$$

$$E^{(i)} = (1 - k_i^2)\, E^{(i-1)}$$

$$c_0 = \ln 1 \tag{6}$$

$$y_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} \tag{7}$$

Where$1 \le m \le p$.
The following explanation applies to the stages in Figure 2 above. [17]

- Acoustic noise from the original auditory input can be eliminated by the pre-emphasis method. Noise is interference or interference from the surrounding real hearing object.
- The frame-blocking process divides the acoustic signal into acoustic segments.
- Windowing detects and sorts out the differences in separate acoustic signals.
- Autocorrelation analysis ensures that the signals from each frame are comparable.
- The process of converting the magnitude of the LPC parameters—the logarithmic area ratio, reflection coefficient, and LPC coefficient—into the outcome of autocorrelation analysis is known as LPC analysis.
- Using a Fourier transform, the cepstral analysis converts the LPC parameter into a cepstral coefficient.

II.2 HIDDEN MARKOV MODEL

One of the secret Markov models is the Hidden Markov Standard Model (NHMM) $\{X_t, Y_t\}_{t\in\mathbb{N}}$Y which presupposes that it can be observed as long as its distribution is known to be normal. Unfollowed, the event's cause creates a Markov chain with the state space. The hidden Markov model's primary flaw is that it optimizes the Y observation process's potential function, which $Y_t X_t$ $\{X_t\}_{t\in\mathbb{N}}$ $S_X = \{1,2,\cdots,m\}$. [18][19] [17]

$$L_T(\phi) = P(Y_1 = y_1, Y_2 = y_2, \dots, Y_T = y_T|\phi)$$

$$= \sum_{i_1=1}^{m} \dots \sum_{i_T=1}^{m} \left(\pi_{y_1 i_1} \pi_{y_2 i_2} \dots \pi_{y_T i_T}\right) \times \left(\delta_{i_1} \gamma_{i_1 i_2} \gamma_{i_2 i_3} \dots \gamma_{i_{T-1} i_T}\right)$$

$$= \sum_{i_1=1}^{m} \dots \sum_{i_T=1}^{m} \delta_{i_1} \pi_{y_1 i_1} \prod_{t=2}^{T} \gamma_{i_{t-1} i_t} \pi_{y_t i_t}. \tag{8}$$

The acoustic properties of corona discharge, etc., were extracted in this work using the linear predictive coding (LPC) method. Parameter estimation, including mean, variance, and transition matrix, is performed to maximize the probability function. The Expectation Maximization (EM) algorithm is used iteratively to optimize the probability function in the equation (9). Each algorithm is formulated as follows.$y_1, y_2, \dots, y_T$[19]. [17]

$$\mu_i^{(k+1)} = \frac{\sum_{t=1}^{T} \alpha_t^{(k)}(i) \beta_t^{(k)}(i) y_t}{\sum_{t=1}^{T} \alpha_t^{(k)}(i) \beta_t^{(k)}(i)}, \tag{9}$$

$$\sigma_i^{2(k+1)} = \frac{\sum_{t=1}^{T} \alpha_t^{(k)}(i) \beta_t^{(k)}(i) \left(y_t - \mu_i^{(k+1)}\right)^2}{\sum_{t=1}^{T} \alpha_t^{(k)}(i) \beta_t^{(k)}(i)} \tag{10}$$

$$\gamma_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T-1} \alpha_t^{(k)}(i) \gamma_{ij}^{(k)} P^{(k)}(y_{t+1}|j) \beta_t^{(k)}(j)}{\sum_{t=1}^{T-1} \alpha_t^{(k)}(i) \beta_t^{(k)}(i)} \tag{11}$$

Forward and backward algorithms can be used to calculate recursivelyand Furthermore, these findings are used to divide the corona release phenomenon based on voltage and interference taking into account the maximum probability.$\alpha_t(i) = P(Y_1 = y_1, Y_2 = y_2, \dots, Y_t = y_t, X_t = i|\phi)$ $\beta_t(i|\phi) = P(Y_{t+1} = y_{t+1}, \dots, Y_T = y_T|X_t = i, \phi)$.

II.3 BACKPROPAGATION NEURAL NETWORK (BPNN)

The backpropagation neural network (BPNN), an artificial intelligence technique that uses optimization to reduce the function of the objective—that is, the difference between the experimental and model data—is applied to model data based on experimental data. Figure 4 below shows the BPNN architecture. [20][21]:

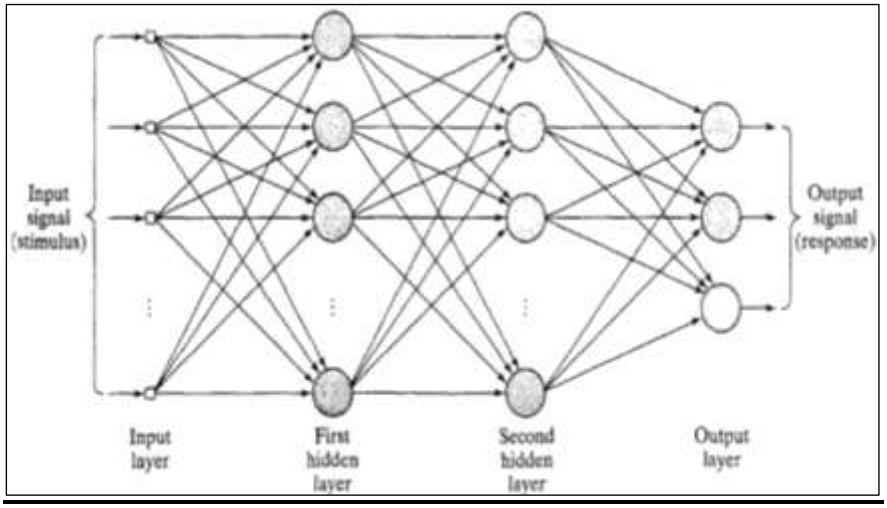


Figure 4: BPNN Architecture.
Source: Authors, (2026).

The objective functions of BPNN are as follows:

$$\xi = \frac{1}{2}\sum_{j\in C} e_j^2(n) \tag{12}$$

with , with is the output signal of BPNN is the experimental output data of the same input. $e_j(n) = d_j(n) - y_j(n) y_j(n)\, d_j(n).$

BPNN learning is achieved by entering the observation data into the BPNN and then reducing errors along with estimating BPNN parameters that match the weight, as shown in Figure 4. To minimize errors in equations (12), various optimization methods are used.

Table 1: Connected functions of the BPNN model.

Connected function	Schema
	BPNN
$z_{in} = f_1 = f_1(X^*, b1, v)$	$b1 + X\ v$
$z = f_2 = f_2(z_{in})$	$\frac{1}{(1 + exp(-z_{in}))}$
$y_{in} = f_3 = f_3(z, b2, w)$	$b2 + z\ w$
$(\alpha, \beta) = f_4 = f_4(y_{in})$	-
$T^* = L = L(\alpha, \beta, C, E, n)$	-
$T^* = f_5 = f_5(y_{in})$	y_{in}
n_1	4
n_2	1

Source: Authors, (2026).

with lifetime input and data$X^*T^{\#}$ [22], the research model for and BPNN in this case is as follows:

$$T^* = f_5\ o\ f_3\ o\ f_2\ o\ f_1 \qquad for\ BPNN \tag{13}$$

In this model, the parameters is b1,v,b2,w for the BPNN scheme.

II.4 PARAMETER MODEL ESTIMATION

To gain a better grasp of the formulation of the model parameters in Table 1, the GD approach is then used to optimize the function based on its objective in equation (13). is the function of the objective model in equation (6), and the model construction is shown in Table 2.T^* Genetic algorithms have phases and function similarly to genes. [23] [24] Initialization, spin wheel, cross-over, mutation, individual assessment, and linear fitness improvement. All of these procedures are executed by default, with the exception of the individual evaluation phase. Change the objective function now in accordance with Table 1 based on [23] [25].

The evolutionary algorithm's stages (1) through (6) guarantee that the BPNN approaches converge, meaning that the model's results get smaller with each iteration, reducing the gap between the experiment and the model. Based on [26][27], to guarantee convergence and minimize errors between models and experimental data. The convergence of the LM technique for BPNN will thus be guaranteed by the modified parameters. Put another way, as the number of iterations increases, the gap between the experiment's and the model's outcomes will narrow. The following table 2 illustrates the entire procedure:

Table 2: Parameters of BPNN Formulation.

Formulation BPNN
$w_{ji}(n+1) = w_{ji}(n) - \alpha e_j(n+1)\, z, \qquad b2_i(n+1) = b2_i(n) - \alpha e_j(n+1),$
$v_j(n+1) = v_j(n) - \alpha e_j(n+1)\left(\frac{1}{(1+exp(-z_{in}))}\frac{exp(-z_{in})}{(1+exp(-z_{in}))}\right) X,$
$b1_j(n+1) = b1_j(n) - \alpha e_j(n+1)\left(\frac{1}{(1+exp(-z_{in}))}\frac{exp(-z_{in})}{(1+exp(-z_{in}))}\right),$
is scalar (learning rate).α

Soucer: [24], [28-31].

III. RESULTS AND DISCUSSION

III.1 DESIGN AND OBSERVATION SERIES

The design and series of observations made can be seen in figures 5 and 6 below.

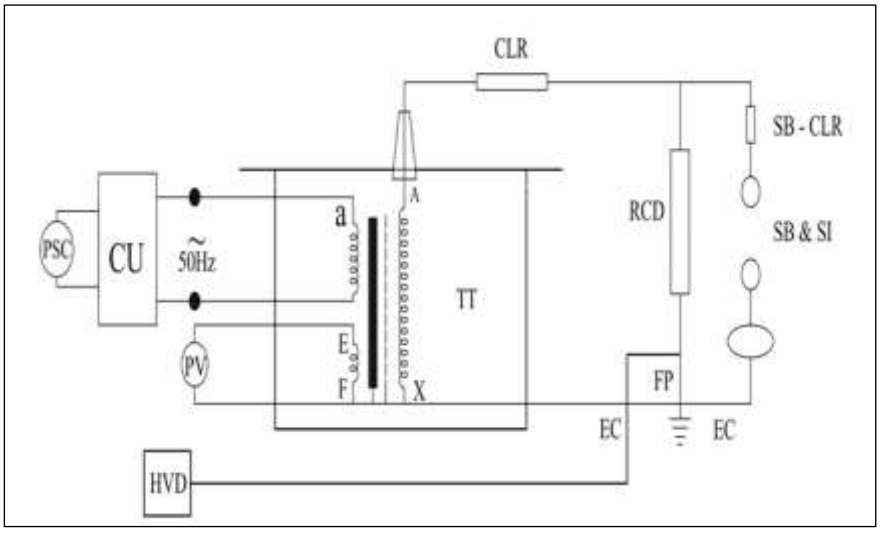


Figure 5: Corona Discharge Acoustic Recording Series in 20 kV Cubic.
Source: Authors, (2026).

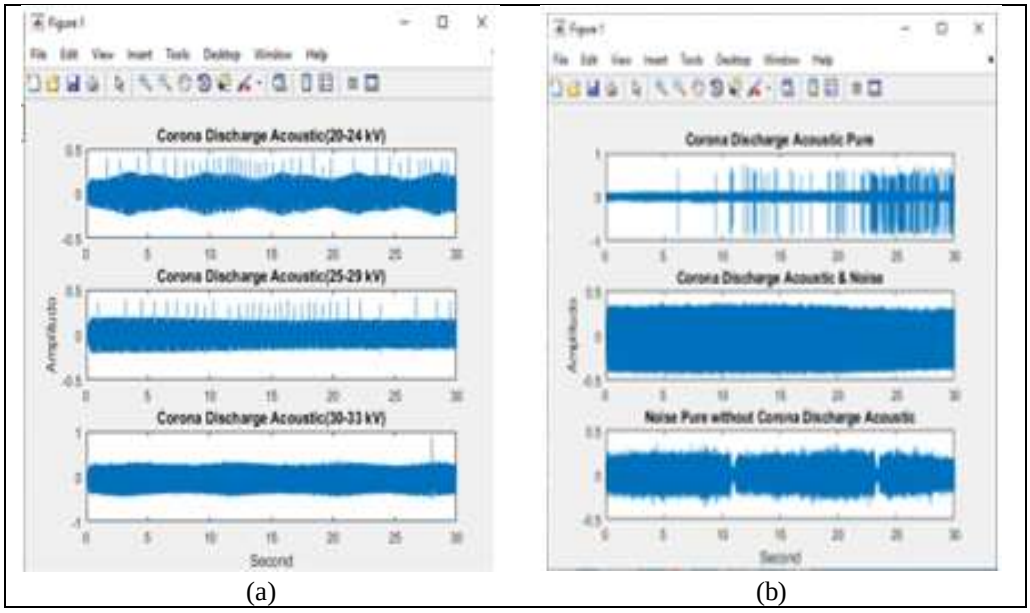


Figure 6: Research Design for Acoustic Recording of Corona Discharge in a 20 kV Cubicle.
Source: Authors, (2026).

Figure 6 displays the acoustic recording of corona discharge in the High Voltage Oratory Lab at the PLN Institute of Technology. It is generated by a test transformer that can be adjusted to 100 kV and has a 200 V scale ratio. According to the findings of the preliminary observations, the corona discharge's penetration voltage was 34.3 kV. In order to create three clusters, the acoustic recording of corona discharge in this investigation started at a voltage of 20 kV and was thereafter increased gradually. Furthermore, three types of CDs were used to capture the acoustic data: pure CDs, CDs with noise, and CDs with pure noise. The temperature in the room increased from 27.5 to 35.3 degrees Celsius over the data collecting period, and the humidity ranged from 70% to 95%. The corona discharge is recorded acoustically between the two electrodes, which are spaced three centimeters apart.[11]. The microphone is positioned 5 cm away from the electrodes for acoustic recording, and the recordings are saved on a computer for analysis.[32]

III.2 ACOUSTIC CORONA EMISSION DATA AND EXTRACTION

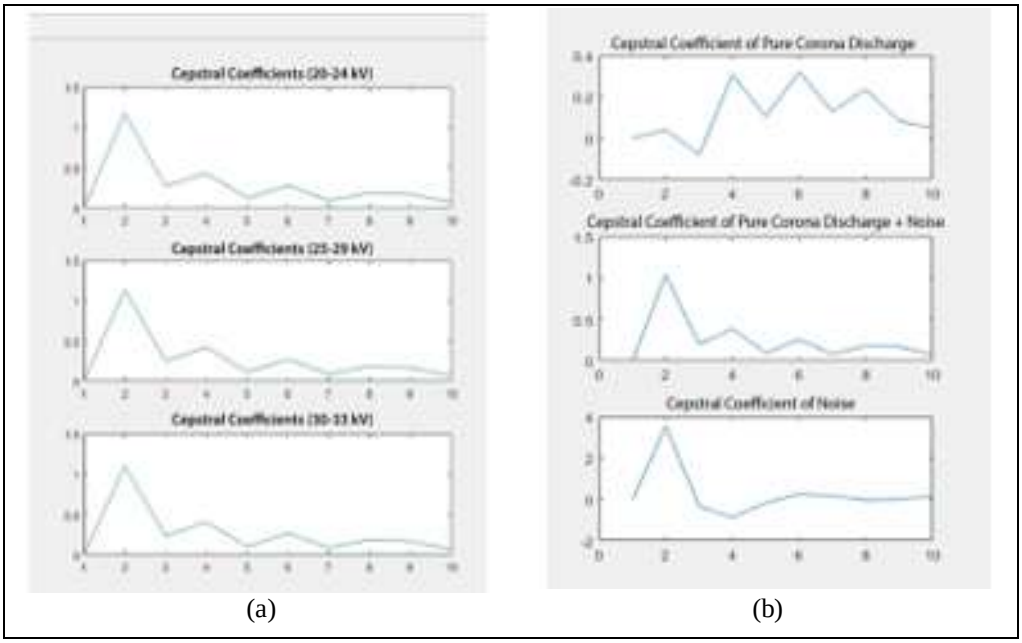
With each pure CD cluster, corona acoustic dissipation was recorded using the research design depicted in figure 5 according to sound and voltage (20–24 kV, 25–29 kV, and 30-33 kV). Figure 6 displays CDs with sound and CDs with only sound. [17]



(a) (b)

Figure 7: Acoustic Coronas Discharge (a) Based on Voltage (b) Based on Noise.
Source: Authors, (2026).

Using equations (1)–(5), the acoustic data in Figures 7 (a) and 6 (b) were calculated to yield the LPC parameters. Moreover, Fourier transformations do a better job of extracting the acquired features.[16]. utilizing equations (6)–(7), which yields Figure 7(a) and 7(b)'s cepstral coefficient of LPC.



(a) (b)

Figure 8: Cepstral Coefficient LPC Acoustic Corona Discharge (a) Based on Voltage (b) Based on Noise.
Source: Authors, (2026).

Figure 8(a) is comparable in that there are no outside disturbances, or noises, so it can only produce sufficient results[17]. In contrast, Figure 8(b) is very different because of the sound effect, so it is expected to produce precise grouping results. Furthermore, the extraction of the cepstral LPC feature was used as an input for the Euclidean, HMM, and BPNN distance methods to identify and classify the corona acoustic discharge.

Table 3: Acoustic Identification Accuracy of CDs by Voltage Using HMM.

NOT.	Training		Testing	
	Amount of Data	Accuracy (%)	Amount of Data	Accuracy (%)
1	1	100	9	88.89
2	2	100	8	83 . 33
3	3	100	7	80 . 95
4	4	100	6	77 . 78
5	5	100	5	73. 33
6	6	100	4	91. 67
7	7	100	3	88,89
8	8	100	2	100
9	9	100	1	100
Average	**100%**		**84,44%**	

Source: Authors, (2026).

Table 3 demonstrates how to use an HMM to identify acoustic CDs based on voltage. They achieved 100% accuracy throughout training and 84.44% accuracy during testing.

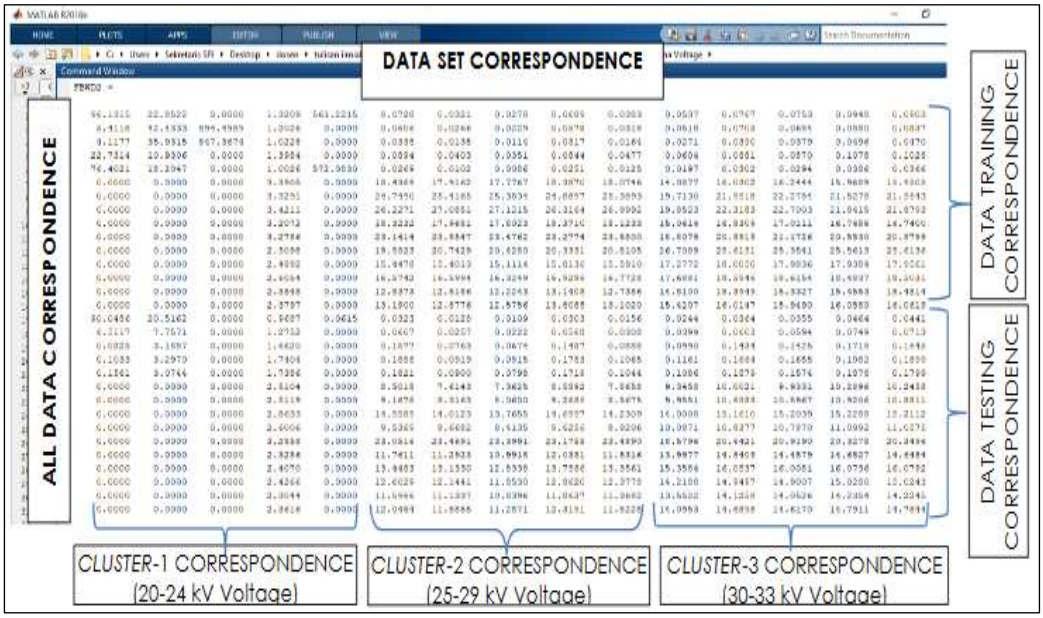


Figure 9: dentification of CD Acoustic Based on Voltage using hidden Markov models.
Source: Authors, (2026).

Table 4: Acoustic Identification Accuracy of CDs Based on Noise Using HMM.

NOT.	Training		Testing	
	Amount of Data	Accuracy (%)	Amount of Data	Accuracy (%)
1	1	100	9	100
2	2	100	8	100
3	3	100	7	100
4	4	100	6	100
5	5	100	5	100
6	6	100	4	100
7	7	100	3	100
8	8	100	2	100
9	9	100	1	100
Average	**100%**		**100%**	

Source: Authors, (2026).

Table 4 demonstrates how to use an HMM to identify CDs acoustically based on noise. They achieved 100% accuracy for both testing and training.

Table 5: Acoustic Identification Accuracy of CDs by Voltage Using BPNN.

NOT.	Training		Testing	
	Amount of Data	Accuracy (%)	Amount of Data	Accuracy (%)
1	1	99.99	9	95.00
2	2	90.00	8	65.55
3	3	98.50	7	90.20
4	4	91.10	6	70.00
5	5	99.05	5	82.35
6	6	93.33	4	98.00
7	7	97.77	3	75.10
8	8	95.63	2	88.95
9	9	98.00	1	76.00
Average	95.93%		82.35%	

Source: Authors, (2026).

Table 5 demonstrates how to use an BPNN to identify acoustic CDs based on voltage. They achieved 95.93% accuracy throughout training and 82.35% accuracy during testing.

Table 6: Acoustic Identification Accuracy of CDs by Noise Using BPNN.

NOT.	Training		Testing	
	Amount of Data	Accuracy (%)	Amount of Data	Accuracy (%)
1	1	98.90	9	100
2	2	90.50	8	100
3	3	99.50	7	100
4	4	91.11	6	100
5	5	99.99	5	100
6	6	92.20	4	100
7	7	98.80	3	100
8	8	93.00	2	100
9	9	97.21	1	100
Average	95.69%		100%	

Source: Authors, (2026).

Table 6 demonstrates how to use an BPNN to identify CDs acoustically based on noise. They achieved 95.69% accuracy for both testing and and 100% accuracy during testing.

Table 7: Recapitulation of CD Acoustic Base Voltage and Noise Using HMM and BPNN.

NOT.	**Method**	**Accuracy Mean (%)**			
		Base Voltage		**Base Noise**	
		Training	Testing	Training	Testing
1	HMM	100	84.44	100	100
2	BPNN	95.93	82.35	95.69	100

Source: Authors, (2026).

Table 7 presents a comparison of the accuracy of the HMM and BPNN methods in classifying Base Voltage and Base Noise conditions in CD acoustic data. HMM shows very high accuracy, particularly for Base Noise, achieving 100% accuracy in both training and testing, while its testing accuracy for Base Voltage reaches 84.44%. BPNN also demonstrates good performance, with training and testing accuracies of 95.93% and 82.35% respectively for Base Voltage, and stable results for Base Noise with 95.69% during training and 100% during testing. Overall, HMM appears more consistent, especially for the Base Noise parameter, although BPNN remains competitive for both parameters.

IV. CONCLUSIONS

This study evaluates the performance of the Hidden Markov Model (HMM) and Backpropagation Neural Network (BPNN) in detecting corona discharge phenomena in 20 kV cubicles using acoustic signals and voltage-based clustering. The application of Linear Predictive Coding (LPC) successfully extracts cepstral features from the recorded corona discharge sounds, enabling effective classification. The results show that HMM achieves outstanding performance for noise-based clustering, reaching 100% accuracy in both training and testing, while its voltage-based testing accuracy reaches 84.44%. In contrast, BPNN provides more balanced and stable results across all categories, with accuracies of 95.93% (training) and 82.35% (testing) for voltage-based clustering, and 95.69% (training) and 100% (testing) for noise-based clustering. These findings highlight that although HMM excels in noise classification, BPNN offers robust performance and better generalization capability for both voltage and noise conditions. Therefore, BPNN is recommended as a reliable early warning method for corona discharge detection in 20 kV cubicles, contributing to improved monitoring, preventive maintenance, and overall distribution network reliability.

V. AUTHOR'S CONTRIBUTION

The authors assisted in the development of a system that detects corona discharge symptoms in 20 kV cubicles using artificial intelligence methods. The authors of this study compared Hidden Markov Model (HMM) and Backpropagation Neural Network (BPNN) techniques to identify sounds resulting from corona discharge symptoms. The authors not only conducted experiments in a high-voltage laboratory to collect sound data, but they also processed the data using the Linear Predictive Coding (LPC) method to identify sound characteristics. The results showed that BPNN can improve identification accuracy compared to HMM. Consequently, BPNN can be used as a more accurate early warning system. This is very important because PLN and related agencies can help detect damage to cubicles quickly, prevent further damage to equipment and improve the reliability of electricity distribution.

VI. ACKNOWLEDGMENTS

The authors would like to thank the PLN Institute of Technology, Jakarta, especially the High Voltage Laboratory, for providing the facilities and opportunity to conduct this research. Appreciation is also extended to fellow lecturers and researchers who assisted in the data collection process and provided valuable input throughout the research.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Christiono Christiono, Miftahul Fikri and Syamsir Abduh.
Methodology: Christiono Christiono, Miftahul Fikri and Syamsir Abduh.
Investigation: Christiono Christiono, Miftahul Fikri and Syamsir Abduh.
Discussion of results: Christiono Christiono, Miftahul Fikri and Syamsir Abduh.
Writing – Original Draft: Christiono Christiono, Miftahul Fikri and Syamsir Abduh.
Writing – Review and Editing: Christiono Christiono, Miftahul Fikri and Syamsir Abduh.
Resources: Christiono Christiono, Miftahul Fikri and Syamsir Abduh.
Supervision: Christiono Christiono, Miftahul Fikri and Syamsir Abduh.
Approval of the final text: Christiono Christiono, Miftahul Fikri and Syamsir Abduh.

VIII. REFERENCES

[1] I. E. Portuguese et al., "RF-based partial discharge early warning system for air-insulated substations," IEEE Trans. Power Deliv., vol. 24, no. 1, pp. 20–29, 2009, doi: 10.1109/TPWRD.2008.2005464.

[2] X. Wang, N. Taylor, and H. Edin, "Effect of humidity on partial discharge in a metal-dielectric air gap on machine insulation and trapezoidal testing voltages," J. Electrostat., vol. 83, pp. 88–96, 2016, doi: 10.1016/j.elstat.2016.08.003.

[3] R. Syahbana, "Analysis of the Formation of Corona in a 20Kv Voltage Cubicle Channel and Its Effect on Power Losses," Lens, vol. 2, no. 48, pp. 14–21, 2019.

[4] R. Masarrang, Lily Stiowaty Patras, and Hans Tumaliang, "Rudolfus Masarrang," J. Tek. Electrical and Computer., vol. 8, no. 2, pp. 67–74, 2019.

[5] and S. K. A. S. Menesy, X. Jiang, M. A. Ali, H. M. Sultan, N. M. Alfakih, "Partial Discharge and Breakdown Characteristics in Small Air Gap Length Under DC Voltage in Needle-Plane Electrode Configuration," pp. 869–874, 2020.

[6] X. Li, J. Wang, T. Lu, and X. Cui, "Statistical analysis of audible noise generated by AC corona discharge from single corona sources," High Volt., vol. 3, no. 3, pp. 207–216, 2018, doi: 10.1049/hve.2017.0159.

[7] A. Syakur and U. Diponegoro, "Measurement of Air Dielectric Penetration Voltage in Various Cells," J. Thesis, no. January 2011, 2016.

[8] C. Widyastuti, I. N. Bagus, and Y. Dharma, "Energy and Electricity: A Scientific Journal of the Impact of Corona on 500 kV SUTET on Radio Interference of Energy and Electricity: A Scientific Journal," J. Ilm., vol. 11, no. 2, pp.
87–97, 2019.

[9] S. Hedtke, M. Pfeiffer, and C. M. Franck, "Corona discharge pulse pattern and audible noise on hybrid AC/DC transmission lines under electric field bias, ripple and ion coupling," J. Electrostat., vol. 102, 2019, doi:
10.1016/j.elstat.2019.103373.

[10] A. J. Moore, M. Schubert, and N. Rymer, "Technologies and Operations for High Voltage Corona Detection with UAVs," IEEE Power Energy Soc. Gen. Meet., vol. 2018-Aug, 2018, doi: 10.1109/PESGM.2018.8585759.

[11] M. Karimi, M. Majidi, H. Mirsaeedi, M. M. Arefi, and M. Oskuoee, "A novel application of deep belief networks in learning partial discharge patterns for classifying corona, surface, and internal discharges," IEEE Trans. Ind. Electron., vol. 67, no. 4, pp. 3277–3287, 2020, doi: 10.1109/TIE.2019.2908580.

[12] M. Wahyudi, Tumiran, I. M. Yulistya Negara, N. Akhmad Setiawan, and B. Sugiyantoro, "Audiosonic Acoustic Detection of Air Corona Discharge based on Fast Fourier Transform," Proc. 2nd Int. Conf. High Volt. Eng. Power Syst. Towar. Sustain. Reliab. Power Deliv. ICHVEPS 2019Pp. 1–6, 2019, doi: 10.1109/ICHVEPS47643.2019.9011029.

[13] M. FIKRI, C. CHRISTIONO, and I. G. MULYANA K., "Clustering of Corona Discharge Phenomena based on Sound using LPC and Euclidean Distance Methods," ELKOMIKA J. Tek. Electric Energy. Tech. Telecommunications. Tech. Electron., vol. 10, no. 3, p. 689, 2022, doi: 10.26760/elkomika.v10i3.689.

[14] N. PASRA, M. FIKRI, K. T. MAURIRAYA, T. RIJANTO, and I. G. P. A. BUDITJAHJANTO, "Noise Detection of Corona Discharge Sound based on Noise using LPC and Euclidean Distance Methods," ELKOMIKA J. Tek. Electric Energy. Tech. Telecommunications. Tech. Electron., vol. 11, no. 1, p. 72, 2023, doi: 10.26760/elkomika.v11i1.72.

[15] Mr. Fikri et al., "Optimization Objective Function Corona Discharge Acoustic Using Fuzzy C-Means (FcM)," Elkha, vol. 15, no. 2, p. 84, 2023, doi: 10.26418/elkha.v15i2.63601.

[16] Mr. FIKRI et al., "Clustering of Corona Discharge Sounds by Voltage using the Fuzzy C-Mean Method," ELKOMIKA J. Tek. Electric Energy. Tech. Telecommunications. Tech. Electron., vol. 11, no. 3, p. 609, 2023, doi: 10.26760/elkomika.v11i3.609.

[17] Mr. Fikri et al., "COMPARISON OF CORONA DISCHARGE IDENTIFICATION IN 20 kV CUBICLES BASED ON VOLTAGE AND NOISE USING ED, HMM, AND FCM," J. Techno., vol. 86, no. 5, pp. 11–22, 2024, doi: 10.11113/jurnalteknologi.v86.19986.

[18] M. FIKRI, B. SETIAWATY, and I. G. P. PURNABA, "Parameter Estimation and Congenescence of Parameter Estimation of the Poisson Hidden Markov Model," J. Math. Its Appl., vol. 15, no. 1, pp. 45–54, 2016, doi: 10.29244/jmap.15.1.45-54.

[19] M. Fikri, Samsurizal, Christiono, and K. T. M, "Weather Modeling Using the Hidden Markov Model for Solar Energy Utilization," Lightning, vol. 9, no. 2, pp. 217–224, 2020.

[20] M. D. Engelhart and H. Moughamian, "Book Reviews: Book Reviews," Educ. Psychol. Stuff., vol. 28, no. 3, pp. 951–951, 1968, doi: 10.1177/001316446802800332.

[21] S. Haykin, Neural Networks - A Comprehensive Foundation, Second. Pearson Prentice Hall, 1988.

[22] Z. Ma, L. Yang, H. Bian, M. S. Bhutta, and P. Xu, "An Improved IPM for Life Estimation of XLPE under DC Stress Accounting for Space-Charge Effects," IEEE Access, vol. 7, pp. 157892–157901, 2019, doi: 10.1109/ACCESS.2019.2946521.

[23] Mr. Melanie, An Introduction to Genetic Algorithms. Cambridge: MIT Press, 1996.

[24] Mr. Fikri et al., "Lifetime estimation of DC XLPE cable insulation using BPNN-IPM improved with various schemes and optimization methods," Indonesian. J. Electr. Eng. Comput. Sci., vol. 36, no. 1, pp. 86–98, 2024, doi: 10.11591/ijeecs.v36.i1.pp86-98.

[25] S. Mirjalili, Evolutionary Algorithms and Neural Networks. Cham: Springer Nature, 2019.

[26] H. B. Nielsen, "Imm Department of Mathematical Modelling Damping Parameter in Marquardt'S Method," no. M, 1999.

[27] C. Kanzow, N. Yamashita, and M. Fukushima, "Levenberg-Marquardt methods with strong local convergence properties for solving nonlinear equations with convex constraints," J. Comput. Appl. Math., vol. 173, no. 2, pp. 321–343, 2005,
doi: 10.1016/j.cam.2004.03.015.

[28] R. P. Lippmann, "An introduction to computing with neural nets," ACM SIGARCH Comput. Archit. News, vol. 16, no. 1, pp. 7–25, 1988, doi: 10.1145/44571.44572.

[29] S. Kusumadewi and S. Hartati, NEURO-FUZZY Integration of Fuzzy Systems and Neural Networks. Yogyakarta: Graha Ilmu, 2006.

[30] R. P. Lippmann, "An Introduction to Computing with Neural Nets," IEEE ASSP Mag.Pp. 4–22, 1987.

[31] D. E. Rumelhart and G. E. Hintont, "Learning representations by back-propagating errors," Nature, vol. 323, no. 9, pp. 533–536, 1986.

[32] Y. Liu, J. Xv, Y. Liu, H. Yuan, and Y. Cui, "A Method for the Indirect Detection of Audible Noise from High-Voltage Direct Current Transmission Lines," IEEE Trans. Instrument. Stuff., vol. 69, no. 7, pp. 4358–4369, 2020, doi: 10.1109/TIM.2019.2942251.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 35-42. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2684

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

DESIGN AN INTELLIGENT CONTROLLER FOR LOAD BALANCING IN WIND ENERGY SYSTEMS CONNECTED TO THE GRID

Zainab Jamal Mohammed*[1] and Mafaz Yahya Damarah[2]

[1]University of Al-Hamdaniya, Nineveh, Iraq.
[2]Dep. Of Electrical technologies, College of Polytechnic, Northern Technical University, Mosul, Iraq.

[1]https://orcid.org/0000-0001-8076-3128, [2]https://orcid.org/0009-0004-0190-0577

Email: *zainab.j.alohmayed1990@uohamdaniya.edu.iq, mti.lec126.mafaz@ntu.edu.iq

ARTICLE INFO

Article History
Received: September 7, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Intelligent controller,
wind energy systems,
on-grid,
electrical power.

ABSTRACT

With the rapid expansion of wind energy integration into electrical grids, there is an urgent need for an intelligent control system capable of adapting to the sharp power fluctuations caused by the ever-changing nature of wind and maintaining stable frequency and voltage to meet fluctuating load requirements. This paper aims to provide complete and intelligent control of a combined wind turbine-to-grid power generation system. The intelligent control unit can instantly estimate wind speed and rated power, thereby providing automatic adjustment mechanisms for inverters and load distribution. The system is represented by a detailed simulation in MATLAB/Simulink. The control unit can also analyze instantaneous frequency and voltage deviations, classify load conditions, and compensate for deficits or surpluses in the wind turbine's generated power. This study evaluated the typical performance of the proposed unit for various scenarios, including unexpected fluctuations in wind speed and sudden load changes. Simulation results showed that the intelligent controller eliminated a significant proportion of frequency deviations compared to conventional control systems. Furthermore, the voltage drops and rises were within acceptable specifications.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Wind turbines hold a leading position in the context of environmental sustainability and meeting the growing global demand for energy. They are among the most prominent renewable energy technologies. They are used to convert wind energy into electrical power. Wind turbines can be classified based on the amount of energy derived from capturing the wind to operate them [1][2]. Wind turbines reduce dependence on fossil fuels, leading to reduced carbon emissions and combating climate change [3]. Wind turbines increase the efficiency of the electrical system when operated in parallel with the main grid [4]. Furthermore, the generated wind energy is transmitted to the electrical grid, providing greater stability to the power system, as wind turbines can supply electricity when demand is high. However, renewable energy sources, such as wind turbines, face challenges including intermittency and fluctuation due to varying wind speeds. Therefore, it is essential to develop strategies for frequency control, grid stability, and load control.

Smart control technologies can also improve turbine performance by enabling them to respond more quickly to changes in energy demand and handle disturbances such as overloads, faults, or even load shortages. With wind power having grown by an exceptional 14% since 2010, it is the second-fastest-growing new energy source after solar power. Wind power is set to become the third-largest source of global electricity generation after hydropower and solar by 2030, according to estimates by the International Energy Agency [5]. This growth is part of the global shift from using polluting resources as fuels to more efficient and environmentally friendly sources. All of these factors combined demonstrate the importance of wind power in the future energy mix. These forecasts point to profound changes coming to the energy industry as countries seek to reduce their dependence on fossil fuels and transition to a cleaner mix of sources. Improving the efficiency of wind turbines and related technological solutions enhances their ability to compete in the global energy market [6][7].

A connection of a wind power generation system (WPS) with the electrical primary grid is referred to as a GRID-on connection. Grid-tied connectivity refers to how a WPS integrates with load-only devices or an energy storage system. When connected to the grid, it needs to meet synchronisation conditions because the three-phase inverter (VSI) voltage and frequency must be equal to the grid's voltage, since that remains constant. On the other hand, the voltage and frequency of the WECS change with wind speed, which requires the development of a robust and reliable controller to reach synchronisation. The voltage at the DC link should be stable; this is the main goal of grid-side control, and we also need monitoring of DC link input voltage to VSI & power transfer active/reactive power into the grid. To achieve this, there is a need to use advanced control technologies to guarantee the stability of the electrical system and an even better plant performance.

These results are important because the wind power generation system can successfully help to achieve an efficient development of energy budgets, and at the same time ensure the stability in the electric system [8]. All these features and more can be achieved by using a three-level inverter as an interface to the grid, ensuring the quality of the output signals and, in general, the durability and responsiveness of the entire system [9]. Given the importance of maintaining constant voltage and frequency, in addition to all the other advantages of connecting renewable energy systems to the grid, many advanced algorithms are emerging to achieve this. In this paper [10], the researcher used the Wind Driven Optimization Algorithm and the Firefly Algorithm together to balance loads and maintain system stability, in addition to comparing them with another group of algorithms. All of this opens up future horizons for hybrid systems and smart cities. While this study deals with a proposal for a fuzzy logic algorithm to track the maximum point of a hybrid renewable energy system consisting of a wind turbine and solar cells, in addition to smart control units capable of managing the power generated by the system and distributing it between the grid or charging batteries[11].

The use of advanced quasi-ZSI (Z-Source Inverters) technology has a significant impact on the stability of both frequency and voltage, achieving high efficiency and reducing the number of components used, unlike conventional designs, in addition to the overall conversion efficiency. Furthermore, it is possible to connect renewable energy sources directly to the grid without the need for additional conversion stages [12]. The system used in this study combines wind turbines and solar cells with a set of control units. Some of these units work to draw the maximum power generated from the system, defined as MPPT algorithms. Other units control the voltage and frequency and maintain them constant in three scenarios discussed: changing radiation intensity, changing wind speed, and load disturbances using a load-side converter (LSC) control unit [13]. The use of smart control methods and artificial intelligence strategies in renewable energy power generation systems has become a must.

Therefore, an infinite number of these strategies have emerged, all of which serve one goal: increasing the capabilities of prediction and adaptation to overcome all obstacles and achieve a faster and more accurate response for the entire system [14]. After reviewing previous research and methodologies, this research proposes a smart FPI controller that controls a VSI on the grid side to synchronize the electrical power generated by wind turbines with the grid in terms of voltage and frequency, as well as balance the supplied loads and handle additional loads treated as external disturbances. All results of the proposed controller will be compared with those of a conventional PI controller. The proposed system model comprises a wind turbine, various other components, and complex control processes, which will not be the focus of this study. This study will, however, cover an intelligent controller that controls the VSI, an isolation transformer, the grid, and loads. Figure 1 shows the system proposed in this study.

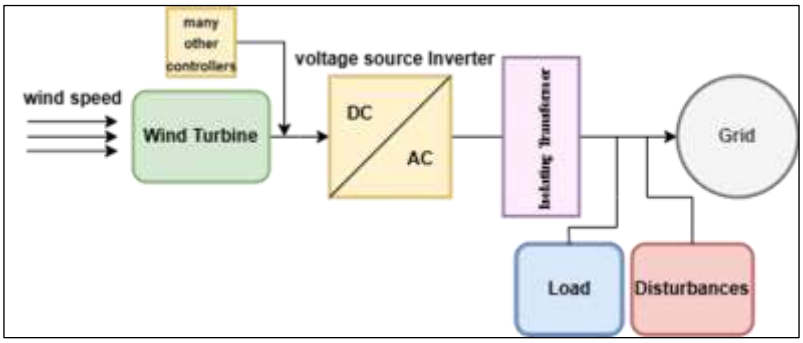


Figure 1: WECS scheme connected to the main grid and equipped with an intelligent controller.
Source: Authors, (2026).

II. THEORETICAL REFERENCE

II.1 WIND TURBINE

The power output of wind turbines varies depending on wind speed, with each wind turbine having its power performance curve. This curve enables the prediction of a wind turbine's power production without considering the technical details of its various components[15]. By analysing performance curves, operators and planners can estimate turbine efficiency under varying wind conditions, thereby contributing to improved energy management strategies and more informed electrical grid planning. Momentum theory is used to analyse the behaviour of wind turbines and make predictions about their performance. In this context, variables are analysed to ensure they have the same value over a specific portion of the air stream pipe, assuming that air is incompressible and fluid motion is constant. The resulting force is expressed by the flow of air, which is moving at a speed. The following equation expresses the mechanical power captured by a wind turbine [16][17]:

The mechanical power captured by a wind turbine

$$P_m = 0.5\rho A v^3 C_p(\beta, \lambda) \quad (1)$$

Where:

ρ = air density (Kg/m^3)
v = wind speed (m/s)
$A = \pi r^2$ (m^2)

The Betz coefficient represents the maximum theoretical value of C_p where:

$$C_{pmax} = 0.593 = 59.3\% \quad (2)$$

II.1.1 Intelligent Controller

A power converter is known as a device that converts DC power into AC power. This device consists of six power switches that utilise IGBT technology with anti-feedback diodes. These keys are divided into two groups: the upper group, which includes keys S1, S2, and S3, and the lower group, which consists of keys S4, S5, and S6[18]. To avoid short circuits, the switching technology used must be intelligent so that two switches on the same terminal do not operate at the same time. As one of the most common methods of switching signal generation, PWM is increasingly used in VSI because it controls the voltage and suppresses the third harmonic of the inverter's output voltage, improving the quality of the generated power [19]. PI controllers remain the most widely used controllers in the wind energy conversion system (WECS) industry due to their robustness and high stability margins. However, to achieve optimal performance from these controllers, tuning key parameters is essential to ensure correct operation and reliable control of the system [20].

Tuning PI controller parameters is a significant challenge in nonlinear or complex systems, where even minor adjustments are difficult for these controllers to handle. In recent years, PI controllers have been replaced by fractional proportional integral (FPI) controllers as a means of overcoming the problems described above. Additional benefits of FPI controllers are based on their ability to adapt to dynamic changes within the system, which in turn makes control more effective and better responsive to changes. The FPI controller is the next generation of PI controllers, offering excellent system performance. FPI controllers have the benefits of increased system stability and reduced sensitivity to parameter changes, making them more suitable for dynamic systems. Moreover, FPI controllers respond faster to reach zero crossing time to dampen oscillation, which directly helps increase the efficiency and stability of the entire system [2][21]. Figure 2 shows how the smart control unit works to enable the effective coordination and work synchronisation between the grid and the wind generator.

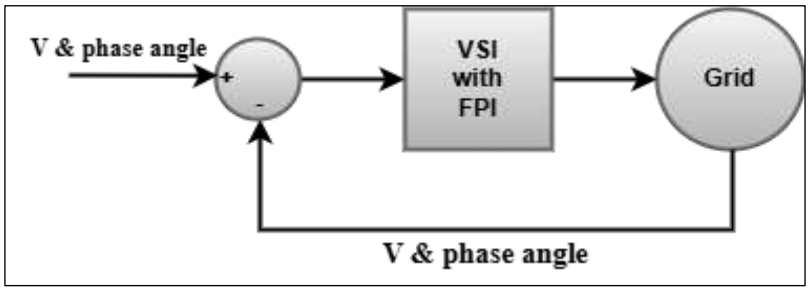


Figure 2: Operation of the intelligent control unit.
Source: Authors, (2026).

Figure 3 demonstrates the load balance between the grid and the wind turbines and how the intelligent unit controls the loads. Rated loads are fed from the wind turbines, and any additional load that introduces disturbances is fed from the grid. When there are no loads, all the power produced by the turbines is transmitted to the grid, maintaining the stability of the electrical system and improving energy efficiency.

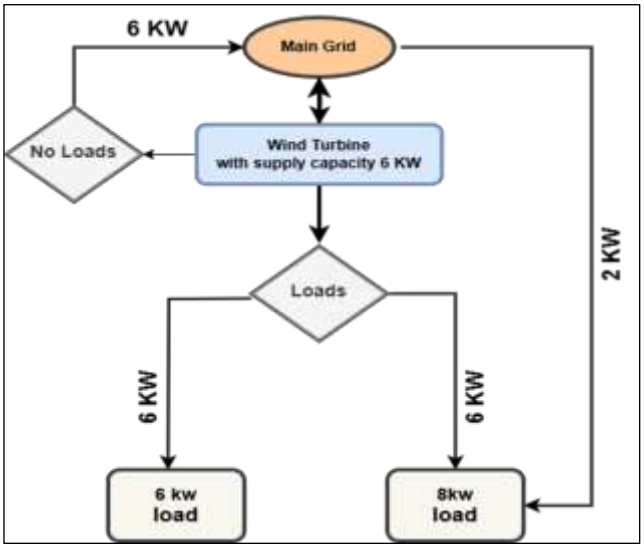


Figure 3: The distribution of loads between the wind turbine and the grid, prioritised according to the diagram, is shown.
Source: Authors, (2026).

II.1.2 Isolating Transformer

Isolation Transformers are utilized to protect sensitive devices from electrical noise and to provide step-up or step-down as required, in addition to coupling two circuits together that cannot be connected directly. These converters remove the DC, while allowing the AC to pass. The proposed three-phase isolation transformer concept is emerging in renewable power system applications due to the improvements in energy transmission and the stability of the electric system [22].

II.1.3 Grid and Loads

The national grid is a complex system designed to transmit and distribute electrical power across large geographic areas, ensuring that consumers' energy needs are met efficiently and reliably. It consists of a group of power plants, transmission lines, and transformer stations, working together to provide electricity to homes, businesses, and industries. Renewable energy sources are often integrated into the national grid, enhancing the sustainability of the electrical system and reducing dependence on fossil fuels. When wind turbines are connected to the grid, they supply the necessary electrical power to the loads. However, it is undeniable that wind turbines alone are insufficient to cover all loads. This is where the intelligent control unit comes in to balance the load distribution between the two sources. Wind turbines supply loads within their rated electrical capacity, while excess or unexpected loads are considered disturbances and are supplied from the main grid, which in turn generates electricity using other sources [23].

III. MATERIALS AND METHODS

This study aims to analyze the performance of a grid-tied wind energy conversion system (WECS) under variable conditions, including changes in load and wind speed, and to demonstrate the adaptability of the control used in the system. During the simulation period ($0 < t < 10$) seconds, the load was constant at 6 kW. Then, in the period ($10 \leq t \leq 12$) seconds, the load was increased by about 34% before returning to its original value. The entire system was modeled using MATLAB, as shown in Figure 4.

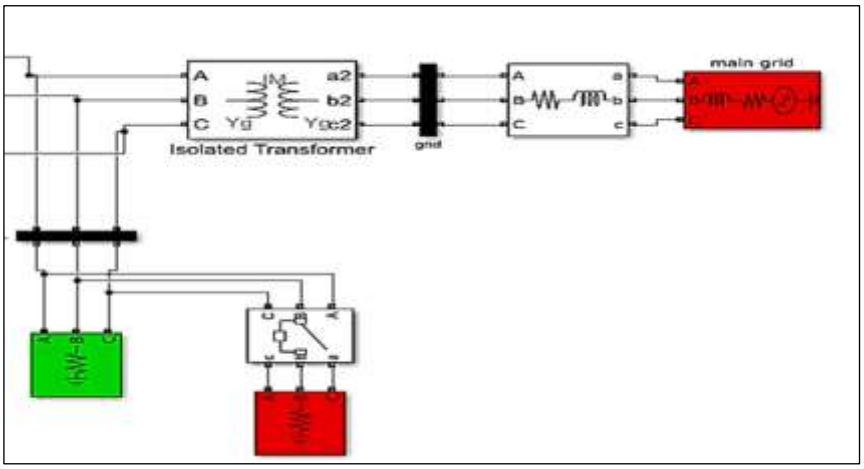


Figure 4: Simulation model of the main grid and loads.
Source: Authors, (2026).

IV. RESULTS AND DISCUSSIONS

The maximum output capacity of the wind turbine is estimated at 6 kW. Therefore, any load exceeding this limit is considered a disturbance and cannot be handled effectively, which is classified as an overload that requires the intervention of additional power sources to ensure the stability of the electrical supply. The wind speed was adjusted according to the step size as shown in Figure 5. This type of wind speed change is considered unrealistic, as it represents a significant challenge for a grid-tied (WECS) due to sudden and rapid transitions between wind speeds. Through this scenario, the effectiveness and efficiency of the proposed adaptive control system can be evaluated.

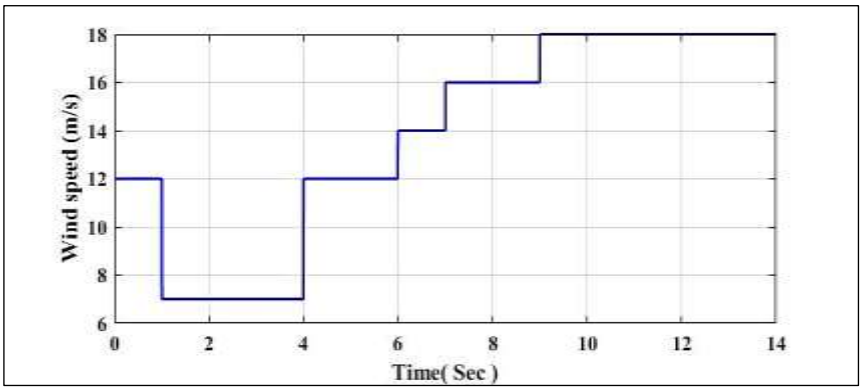


Figure 5: Wind speed variation with time.
Source: Authors, (2026).

Despite fluctuations in wind speed and varying loads, the single-phase voltage waveforms of the load and grid, as shown in Figures 6 and 7, show outstanding performance using both FPI and conventional PI controllers. Voltages remain constant in both methods, but the use of an FPI controller provides better control, as the system exhibits a larger and smoother damped response, reducing oscillations and overshoots in the voltage waveform.

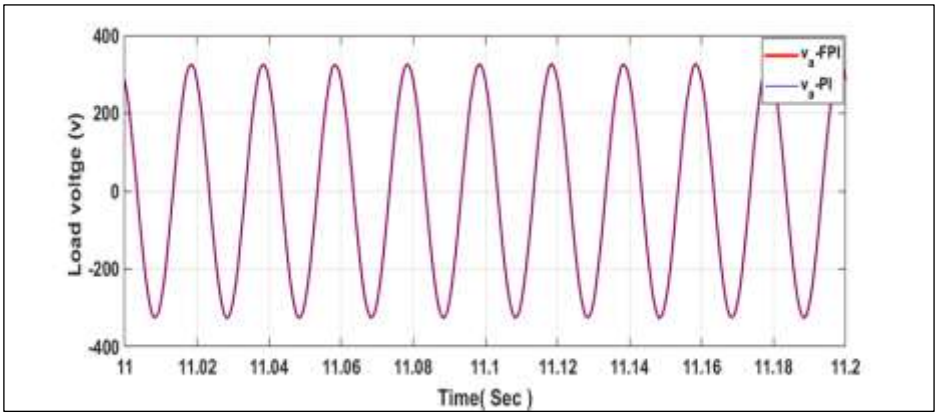


Figure 6: Single-phase load voltage under the intelligent controller.
Source: Authors, (2026).

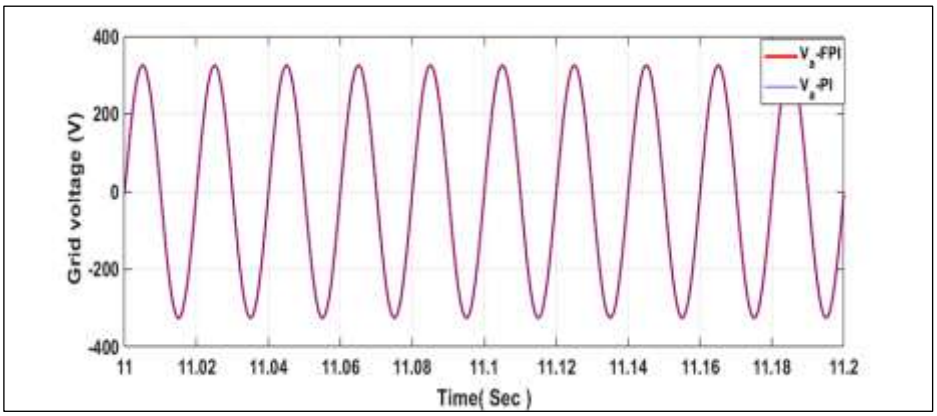


Figure 7: Single-phase grid voltage under the intelligent controller.
Source: Authors, (2026).

Figure 8 shows the single-phase load current. From this figure, it can be seen that the current remains constant at its rated value during the simulation period from (0 to 10) seconds, where the load is 6 kW and is fed entirely from the wind turbine. When the load increases during the simulation period from (10 to 12) seconds to reach 8 kW, this increases the drawing current from the electrical grid, which means that the system draws an additional 2 kW from the grid. The system shows a significant improvement in reducing oscillations and over- and undershoots when using the intelligent controller in the presence of FPI.

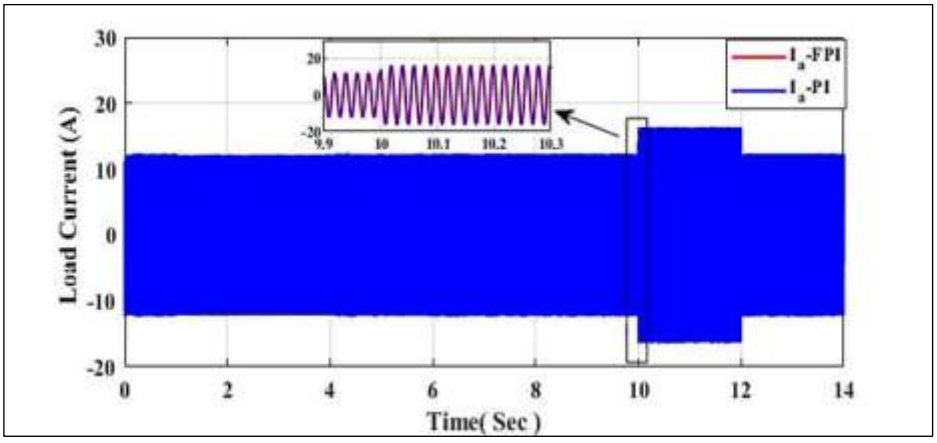


Figure 8: Single-phase load current under the intelligent controller.
Source: Authors, (2026).

Figure 9 shows the single-phase grid current, where it can be seen that the current changes depending on wind speed fluctuations and load changes. Overloads or underloads resulting from low wind speed are always compensated for by the grid, ensuring stability and continuity of electrical supply.

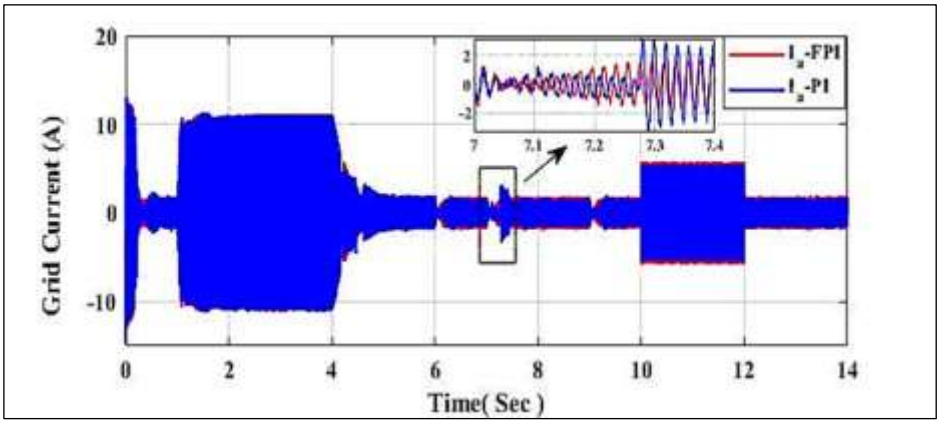


Figure 9: Single-phase grid currents under the intelligent controller.
Source: Authors, (2026).

Using intelligent control technology, the grid frequency is always stable at 50Hz despite abnormal situations. Figure 10 shows the network frequency.

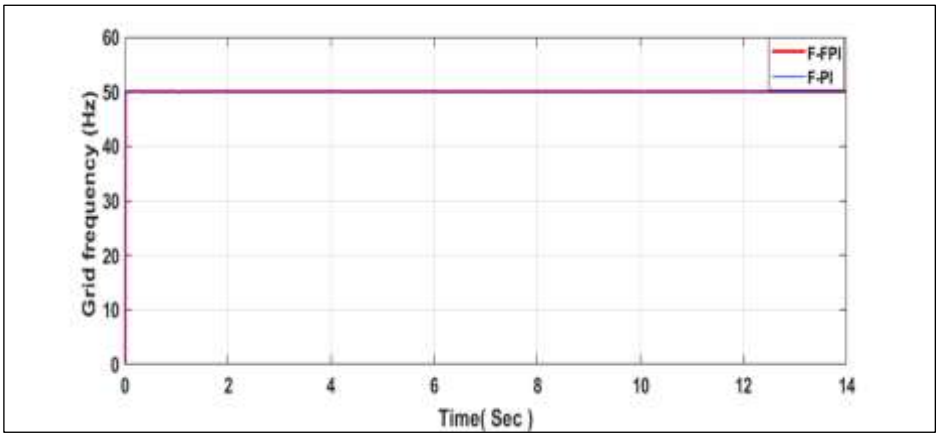


Figure 10: Grid frequency.
Source: Authors, (2026).

Figure 11 shows the power management system under variable wind speed (step size) and variable loads with the intelligent controller, where the system has minimal undershooting and overshooting as well as lower fluctuation.

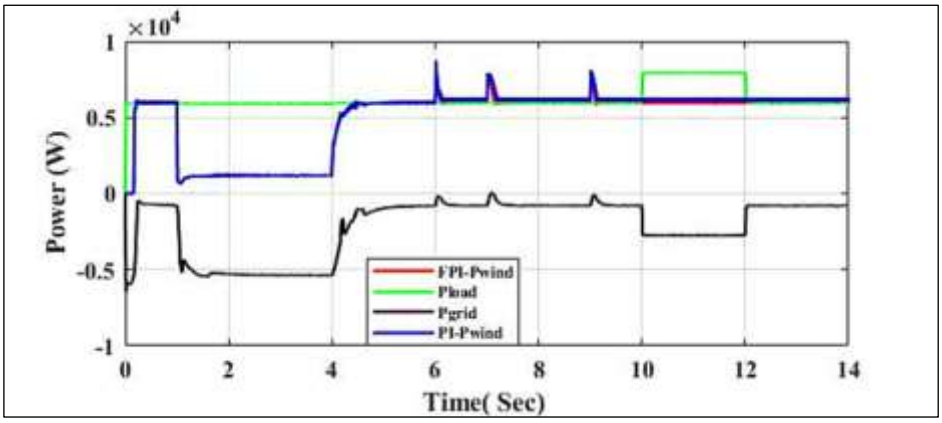


Figure 11: The power management system
Source: Authors, (2026).

V. CONCLUSIONS

This study is a complete evaluation of the performance of grid-tied WECS under variable conditions, including wind speed fluctuations and load variations. An intelligent controller was used with two strategies: A conventional PI controller and an FPI controller. The last one has been used to improve system performance and reduce oscillations and overshoots in the voltage waveform. The results showed that FPI controllers provide better performance in terms of reducing oscillations, which enhances the stability of the electrical system. Load balance has been achieved between the main grid and the wind turbines, as the rated loads are fed from the turbines, while overloads from the grid are compensated. The single-phase load current was also analyzed, and the current remained constant at its rated

value during all times, indicating the effectiveness of the system in meeting the load needs. When the load increased, the draw current from the grid increased, which emphasizes the ability to handle overload. Finally, it is explained that the grid current changes with wind speed fluctuations and load changes, so overloads or lower power generated from the wind turbine are always compensated by the grid, ensuring the stability of the electrical system. These results are evidence of the effectiveness of using the smart controller unit in improving the performance of grid-connected WECS systems, which enhances the reliability of electrical supply and reduces the negative effects of wind speed fluctuations and variable loads, in addition to managing the supply of loads with power in all circumstances.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Zainab Jamal Mohammed and Mafaz Yahya Damarah.
Methodology: Zainab Jamal Mohammed and Mafaz Yahya Damarah.
Investigation: Zainab Jamal Mohammed and Mafaz Yahya Damarah.
Discussion of results: Zainab Jamal Mohammed.
Writing – Original Draft: Zainab Jamal Mohammed.
Writing – Review and Editing: Zainab Jamal Mohammed.
Resources: Zainab Jamal Mohammed and Mafaz Yahya Damarah.
Supervision: Zainab Jamal Mohammed and Mafaz Yahya Damarah.
Approval of the final text: Zainab Jamal Mohammed and Mafaz Yahya Damarah.

VII. REFERENCES

[1] M. H. Soliman, H. E. A. Talaat, and M. A. Attia, "Power system frequency control enhancement by optimization of wind energy control system," Ain Shams Eng. J., vol. 12, no. 4, pp. 3711–3723, 2021, doi: 10.1016/j.asej.2021.03.027.

[2] Z. Jamal Mohammed, S. Enad Mohammed, and M. Obaid Mustafa, "Improving the Performance of Pitch Angle Control of Variable Speed Wind Energy Conversion Systems Using Fractional PI Controller," 2022 Iraqi Int. Conf. Commun. Inf. Technol., pp. 209–215, 2023, doi: 10.1109/iiccit55816.2022.10010639.

[3] Global Wind Energy Council, "GLOBAL WIND REPORT Navigating the global wind power market," vol. Annual Mar, pp. 32–33, 2014, [Online]. Available: http://www.gwec.net/wp-content/uploads/2015/03/GWEC_Global_Wind_2014_Report_LR.pdf

[4] N. W. Miller and J. J. Sanchez-gasca, "Modeling of GE Wind Turbine-Generators for Grid Studies Prepared by :," no. January 2010, 2008.

[5] IRENA, Global Energy Transformation: A Roadmap to 2050 (2019 Edition). 2019.

[6] R. Lacal Arantegui and A. Jäger-Waldau, "Photovoltaics and wind status in the European Union after the Paris Agreement," Renew. Sustain. Energy Rev., vol. 81, no. December 2017, pp. 2460–2471, 2018, doi: 10.1016/j.rser.2017.06.052.

[7] R. K. Behara and A. K. Saha, "Artificial Intelligence Control System Applied in Smart Grid Integrated Doubly Fed Induction Generator-Based Wind Turbine: A Review," Energies, vol. 15, no. 17, 2022, doi: 10.3390/en15176488.

[8] D. Ying, A. Rashid, L. H. Sheng, T. De, T. Shize, and A. Iqbal, "Research on power increase adaptive control strategy based on 5 MW wind turbine," Energy Reports, vol. 7, pp. 50–57, 2021, doi: 10.1016/j.egyr.2021.02.024.

[9] G. A. Adamidis and T. G. Nathenas, "Variable speed wind turbine generator-three level VSI interface," Proc. - 2012 20th Int. Conf. Electr. Mach. ICEM 2012, pp. 2184–2191, 2012, doi: 10.1109/ICElMach.2012.6350185.

[10] T. R. G. Swarna Priya R.M., Sweta Bhattacharya, Praveen Kumar Reddy Maddikunta, Siva Rama Krishnan Somayaji, Kuruva Lakshmanna, Rajesh Kaluri, Aseel Hussien, "Load balancing of energy cloud using wind driven and firefly algorithms in internet of everything," J. Parallel Distrib. Comput., vol. 142, pp. 16–26, 2020.

[11] D. Sarathkumar, M. Srinivasan, A. A. Stonier, R. Samikannu, and D. Vijay Anand, "Design of Intelligent Controller for Hybrid PV/Wind Energy Based Smart Grid for Energy Management Applications," IOP Conf. Ser. Mater. Sci. Eng., vol.
1055, no. 1, p. 012129, 2021, doi: 10.1088/1757-899x/1055/1/012129.

[12] S. S. F. Y, "INTELLIGENT CONTROL OF HYBRID POWER PLANTS WITH WIND TURBINES, PV SOLAR SYSTEMS, AND ENERGY STORAGE SYSTEMS," 2023.

[13] S. Ponnuru, R. A. Kumar, and N. M. J. Swaroopan, "Intelligent control and power management of wind-solar integration of renewable energy sources using microgrid," Mater. Today Proc., vol. 45, no. xxxx, pp. 2323–2328, 2021, doi:
10.1016/j.matpr.2020.10.687.

[14] N. O. Farrar, M. H. Ali, and D. Dasgupta, "Artificial Intelligence and Machine Learning in Grid Connected Wind Turbine Control Systems: A Comprehensive Review," Energies, vol. 16, no. 3, pp. 1–25, 2023, doi: 10.3390/en16031530.

[15] G. Abad, J. López, M. A. Rodríguez, L. Marroyo, and G. Iwanski, Doubly Fed Induction Machine. 2011. doi: 10.1002/9781118104965.

[16] G. S. Kaloi, J. Wang, and M. H. Baloch, "Active and reactive power control of the doubly fed induction generator based on wind energy conversion system," Energy Reports, vol. 2, pp. 194–200, 2016, doi: 10.1016/j.egyr.2016.08.001.

[17] O. Apata and D. T. O. Oyedokun, "An overview of control techniques for wind turbine systems," Sci. African, vol. 10, p. e00566, 2020, doi: 10.1016/j.sciaf.2020.e00566.

[18] S. Vashishtha and K. R. Rekha, "A survey: Space vector PWM (SVPWM) in 3φ voltage source inverter (VSI)," Int. J. Electr. Comput. Eng., vol. 8, no. 1, pp. 11–18, 2018, doi: 10.11591/ijece.v8i1.pp11-18.

[19] N. I. Raju, M. S. Islam, and A. A. Uddin, "Sinusoidal pwm signal generation technique for three phase voltage source inverter with analog circuit & simulation of pwm

inverter for standalone load & micro-grid system," Int. J. Renew.
Energy Res., vol. 3, no. 3, pp. 647–658, 2013.

[20] M. N. Ambia, H. M. Hasanien, A. Al-Durra, and S. M. Muyeen, "Harmony search algorithm-based controller parameters optimization for a distributed-generation system," IEEE Trans. Power Deliv., vol. 30, no. 1, pp. 246–255, 2015, doi:
10.1109/TPWRD.2014.2358940.

[21] Z. J. Mohammed, M. O. Mustafa, and S. E. Mohammed, "Adaptive Hybrid Pitch Angle Control and MPPT for PMSG-Based Wind Power Generator Systems," Int. J. Energy Convers., vol. 10, no. 5, pp. 153–161, 2022, doi:
10.15866/irecon.v10i5.22586.

[22] Z. Cen, "Modeling and Simulation for an 8 kW Three-Phase Grid-Connected Photo-Voltaic Power System," Open Phys., vol. 15, no. 1, pp. 603–612, 2017, doi: 10.1515/phys-2017-0070.

[23] S. Bhatnagar, R. Jangid, and K. Parikh, "Modeling and Design of Maximum Power Point Tracking System Control Algorithm for Pmsg Based Grid Connected Wind Power Generating Unit," Int. J. Tech. Res. Sci., vol. 4, no. 7, pp. 9–20,
2019, doi: 10.30780/ijtrs.v04.i07.002.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 43-58. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2715

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

DARK PATTERNS IN DIGITAL DESIGN: A REVIEW AND ASSESSMENT VIA RISK, TRUST AND DIVERSITY SCORES

Preeti Marwaha[1], Shalu Mahajan[2], Ishitva Joshi[3], Kriti Mishra[4], Ananya[5], Priyanka Sharma[6] and Arunita Chaukiyal[7]

[1, 2, 3, 4, 5, 6, 7]Department of Computer Science, Acharya Narendra Dev College, University of Delhi, New Delhi-110019, India.
[2]Department of Commerce, Acharya Narendra Dev College, University of Delhi, New Delhi-110019, India.

[1]https://orcid.org/0009-0000-7425-4749, [2]https://orcid.org/0009-0004-0591-5995, [3]https://orcid.org/0009-0002-8841-1295
[4]https://orcid.org/0009-0006-7957-3558, [5]https://orcid.org/0009-0005-4234-1174, [6]https://orcid.org/0009-0009-3341-6823
[7]https://orcid.org/0009-0008-5185-5945

Email: preetimarwaha@andc.du.ac.in, shalumahajan@andc.du.ac.in, ishitvajoshi@outlook.com, kritimisra87@gmail.com, itsmeanayasrivastava@gmail.com, priyankasharma@andc.du.ac.in, arunita@andc.du.ac.in

ARTICLE INFO

Article History
Received: September 22, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Dark Patterns,
Deceptive Patterns,
Trust,
Risk,
Diversity.

ABSTRACT

Digital platforms increasingly shape everyday life, yet their design choices often undermine consumer welfare using dark patterns which are deceptive interfaces that manipulate users' decisions. These practices, ranging from hidden costs in e-commerce to manipulative cookie consent prompts and obstructive subscription cancellations, compromise user autonomy and impose financial, privacy, and psychological harms. The risks vary in severity and reversibility, with some patterns leading to lasting detriment while others create cumulative burdens over time. Beyond immediate harm, dark patterns erode trust in digital systems by fostering frustration, skepticism, and disengagement, weakening the long-term relationship between consumers and platforms. Their persistence reflects strong incentives tied to short-term business metrics such as conversions and sign-ups, combined with weak regulatory oversight. Notably, the diversity of dark patterns like urgency cues, obfuscation, misdirection, and coercion etc. illustrates both their adaptability and the challenges of detection. This paper reviews the dark patterns on digital platforms and proposes a structured framework to evaluate dark patterns across three dimensions: risk, trust erosion, and diversity. It also explores mitigation strategies through fair design, transparency, and accountability, while considering the ethical and legal responsibilities of digital platforms. By centering the analysis on consumer trust and risk, the study highlights the pressing need for robust safeguards to counteract diverse manipulative designs and protect users in digital ecosystems.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Trust is a foundational element of user engagement within digital environments, influencing behaviors across e-commerce, social media, mobile applications, and voice interfaces. Users depend on the assumption that digital platforms operate in good faith, offering clarity, protecting user interests, and fostering informed choices [1][2]. This trust is not only technical but also social and psychological, shaped by interface design, past user experiences, and perceived transparency [3]. When systems behave predictably, provide honest feedback, and respect user autonomy, users are more inclined to share data and develop sustained platform relationships [4]. However, this trust is increasingly jeopardized by the widespread use of deceptive interface strategies known as dark patterns, manipulative design tactics that coerce users into unintended actions such as sharing personal data, consenting to cookies, or subscribing without awareness [5]. These practices often emerge when business objectives, such as maximizing engagement or revenue, override ethical concerns [6]. Designers may justify such tactics under commercial pressure, even when user autonomy and welfare are compromised [7]. A common example includes cookie banners that visually emphasize "accept all" options or use misleading phrasing to steer user decisions, undermining genuine consent and eroding trust over time [8]. This erosion can fuel user skepticism, platform backlash, and broader

resistance against the digital economy and its governance mechanisms [9]. Critically, distinguishing between authentic trust and what scholars' term "fake trust", trust that is artificially constructed through manipulation or misrepresentation, is necessary for safeguarding user agency [10]. Fake trust may be built through false information, deceptive designs, or persuasive tactics that blur the line between autonomy and coercion [11]. In response to these challenges, the fields of Human-Computer Interaction (HCI), law, and AI ethics have begun developing systematic frameworks to detect, classify, and regulate dark patterns [12]. This interdisciplinary effort seeks to realign digital systems with principles of fairness, transparency, and user-centered design, advancing legal, technical, and ethical safeguards to ensure digital environments which genuinely respect user autonomy [13][14]. Dark patterns are the interface design strategies that deliberately manipulate, coerce, or mislead users into making choices that are not in their best interest, have become a central concern in the fields of human-computer interaction, design ethics, and digital regulation [15].

Coined by Brignull in 2010, the term originally described deceptive design techniques in user interfaces that were used in specific situations to manipulate user behavior and increase actions like sign-ups, purchases, or other forms of conversion. However, over the past decade, these practices have grown in sophistication and scale, evolving into systematic, pervasive design patterns that exploit cognitive biases and information asymmetries [16]. Today, dark patterns are embedded across a wide range of digital environments, including e-commerce websites that use countdown timers or hidden costs [17], mobile applications that obscure data deletion options [18] , social media platforms that employ infinite scrolls and manipulative notification systems [19], and even newer domains such as voice assistants, digital games, augmented reality (AR), and extended reality (XR), where manipulation can be more immersive and less visible to users [20]. As public awareness and regulatory scrutiny have grown, dark patterns have attracted multidisciplinary scholarly attention, from computer science, law, behavioral psychology, and design ethics. This literature review provides a thematically structured synthesis of the current academic and technical discourse surrounding dark patterns. Specifically, it maps the research landscape across six key domains:

1. **Foundational Work and Taxonomy Development**: These classify dark patterns by intent, technique, and impact, guiding ethical design and regulatory understanding.
2. **Application Domains of Dark Patterns**: Dark patterns are expanding into emerging technologies like voice interfaces, extended reality (XR), and AI-driven systems. These contexts introduce new manipulation risks such as misleading voice prompts, immersive consent traps, or biased AI recommendations, making detection harder. As interfaces evolve, so must ethically design standards to safeguard user autonomy and trust.
3. **Automated Detection of Dark Patterns**: Detection of deceptive design uses machine learning and UI analysis to identify manipulative interface elements. ML models analyze visual layouts, text, and user behavior, while UI analysis applies heuristics and design audits. Together, they reveal patterns that undermine user trust, agency, and experience, supporting efforts toward ethical and transparent design.
4. **Mitigation strategies for dark patterns**: Mitigation strategies for dark patterns include enforcing stricter regulations, ensuring transparent design practices, adopting ethical UX guidelines, and conducting regular audits. Educating users to recognize manipulative tactics, empowering them with clearer choices, and promoting accountability among organizations also reduce risks. Collaboration between policymakers, designers, and watchdogs strengthens long-term protection.
5. **Legal and ethical Analysis**: Deceptive design tricks users, damages trust, limits control over choices, and creates stress, resulting in harmful and unethical digital experiences. These target dark patterns through data protection laws like the General Data Protection Regulation (GDPR) in the European Union, California Consumer Privacy Act (CCPA), and Digital Personal Data Protection Act (DPDP) in India. These frameworks mandate transparency, informed consent, and user control to ensure ethical, user-centric digital experiences and protect privacy.
6. **Methodology for calculating metrics (Risk score, Trust Score and Diversity Score)**: Risk Score evaluates probability and impact of negative outcomes using weighted factors. Trust Score measure's reliability, compliance, and transparency via normalized indicators. Diversity Score quantifies representation across categories using Shannon Entropy.

To trace how Dark Pattern research has evolved, we generated a Storyboard Timeline figure from Scopus bibliometric data (as shown in figure 1 and figure 2) using python. Titles, abstracts, and author keywords were text-mined using curated dictionaries for the major themes: Types, Detection, Mitigation, Psychological Impacts, Ethics/Regulation, and scoring metrics. For each year, we calculated the proportion of publications mentioning these themes and plotted them as continuous lines with annotations that narrate the storyline. The results reveal a clear progression: research initially emphasized defining and categorizing Dark Pattern types, then shifted toward building detection methods, and more recently has begun exploring mitigation strategies. On the psychological–ethical/regulatory–metrics axis, early work focused on user psychology, followed by a growing body of studies addressing ethics and legal regulation, while metrics such as trust, diversity, and risk scores remain only sporadically represented. Taken together, the storyboard timelines show a maturing field that is moving from descriptive taxonomies, through diagnostic tools, toward prescriptive interventions—yet they also highlight mitigation practices and standardized evaluation metrics as key gaps for future research.

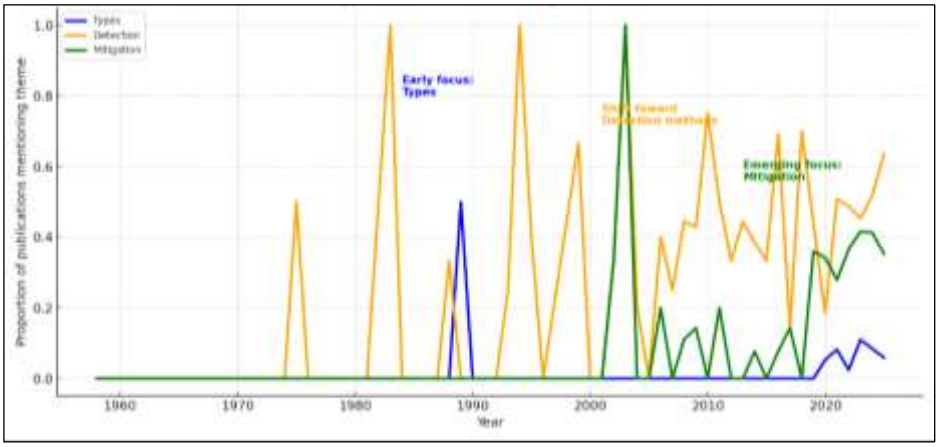


Figure 1: Storyboard Timeline-Types, Detection and Mitigation in Dark Patterns Research.
Source: Authors, (2026).

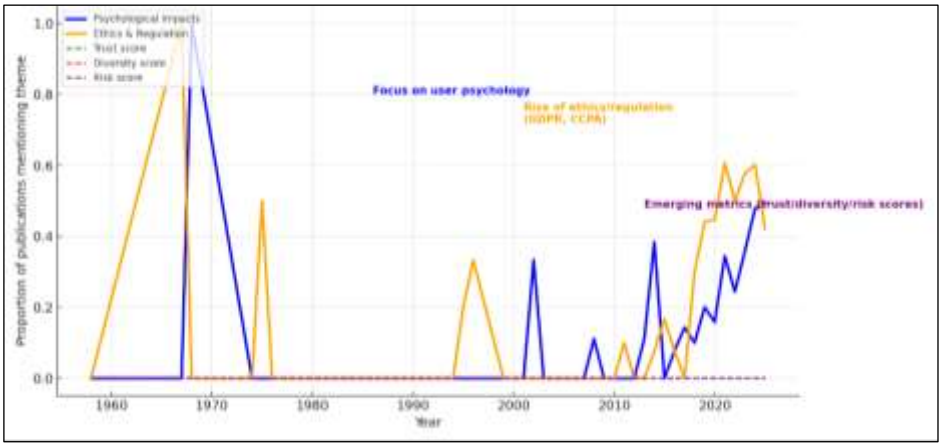


Figure 2: Storyboard Timeline-Psycological, Ethics/Regulation and Metrics in Dark Patterns Research.
Source: Authors, (2026).

II. FOUNDATIONAL WORK AND TAXONOMY DEVELOPMENT

The research around dark patterns revolve around designs that trick users into doing things they might not want to do, like giving up private information or making a purchase. It began with efforts to define and organize them into categories. Two landmark contributions in this area are the works of Gray et al. [5][6] and Gunawan et al. [22]-[21], both of which laid important foundations for understanding and detecting dark patterns in real-world user interfaces. Gray et al. conducted a seminal empirical study that examined real-world user interfaces to identify recurring manipulative design tactics. Their typology was grounded in a qualitative, ethnographic analysis and highlighted the psychological impact of these patterns on user autonomy. Gray et al.'s contribution was foundational, in establishing an ethical lens to assess dark patterns, emphasizing their capacity to erode user trust and agency. Building upon this framework, Gunawan et al. provided one of the most comprehensive classification systems to date. Their research extended the scope beyond web interfaces to include mobile platforms, highlighting the adaptability of dark patterns across different contexts. Their work focused on the structural features of dark patterns and their economic motivations, contributing a more systematic and platform-aware perspective. Zagal et al. [23]-[22] explored how video games use similar tricks like making players spend money to win and showed that these patterns can also appear in entertainment systems. Beyond simply listing and testing these patterns, other researchers have focused on their ethical implications. Devarmani et al. [7] argued that dark patterns aren't just annoying but unfair and morally wrong. They compared them to "nudges" (a term from behavioral science) that push people toward certain decisions in unethical ways. These researchers believe that dark patterns reduce people's freedom to choose and make online experience manipulative, which raises serious concerns about design responsibility and user rights. Table 1 gives the classification of Dark Patterns.

Table 1: Classification of Dark Patterns.

Category	Definition	Techniques / Subtypes	Examples
Information Concealment	Techniques that obscure, hide, or delay the presentation of key information from users, often encourage uninformed decisions.	• Hidden Information • Obfuscation • Alphabet Soup	• Amazon hiding subscription details under interface layers [21]-[23] • Legal/financial information buried in privacy policies [24] • Legalese/acronyms in Term and conditions [5]

Manipulative Wording or Design	Language and visual styles designed to mislead or hide true intent.	• Trick Questions • Misleading Reference Pricing	• Checkbox: "Uncheck this box if you do not want emails" [15] • Inflated original prices to simulate discounts [7]
Choice Architecture Exploitation	Structuring options which manipulate users into making specific, often suboptimal, decisions.	• Preselection • Drip Pricing	• Pre-ticked subscription/ product add-ons [16] • Fees revealed only at final checkout [5]
Coercive Design	Use of emotional pressure or psychological tactics to compel users to act.	• Forced Action • Nagging • Emotional Appeals Urgency	• Provide personal info or accept cookies to proceed [25] • Repetitive exit prompts [26] • "You'll make your cat sad…" [9]. Fake countdown timers [17]
Obstruction	Designs that intentionally make it difficult or time-consuming to take actions against the platform's interests.	• Complete Obstruction • Temporary Obstruction	• Multi-step or call-only cancellations [9] • Privacy settings buried in menus [27]
Sneaking	Actions are performed on behalf of the user without explicit consent or awareness.	• Hidden Costs • Disguised Ads • Drip Pricing • Manipulated Fairness Perception	• Add-ons like insurance, donation preselected [16] • Sponsored content disguised as native posts [5] • Sequential surcharge disclosure [28] • Manipulated Fairness Perception involves designing pricing or presentation tactics that make unfair or inflated costs seem reasonable [29]
Social Proof and Scarcity	Using real or simulated peer activity or limited availability to pressure decision-making.	• Low Stock Claims • Recent Purchase Alerts	• "Only 1 room left!"—often fake [27] • "Maria from London just bought this [17]
Social Engineering	Leverages psychological tricks and peer influence to manipulate behavior.	• Scarcity Claims • Social Proof • Shaming • Countdown Timer • Activity Messages • Fear of Missing Out (FOMO) • Friction in Account Deletion	• Scarcity Claims: "Only 2 left in stock!" to pressure purchases. • Social Proof: "500 people bought this today!" • Shaming: Labels like "No thanks, I prefer to pay full price." • Activity Messages: Fake updates like "Jessica just bought this". • Countdown Timers: Create urgency even if not real.
Gamification	Use of game elements to encourage repetitive use or spending.	• Pay-to-Play • Grinding • Countdown Ads • Infinite Scrolling • Addictive Design • Compulsion Loops • Reduced Friction to Keep Playing/Paying • Artificial Scarcity • Fear of Missing out (FOMO) • Power Imbalance via Purchases • Escalating Rewards • Anticipatory Triggers • Loot Box Rewards • Reward Saturation • Friction in Account Deletion • Bright Patterns Replacement	• Pay-to-Play: Must spend money to progress. • Grinding: Repetitive tasks required for minor rewards. • Countdown Ads: Must watch full ads before proceeding. • Infinite Scrolling: No stopping point to encourage endless use. • Addictive Design: Features like pull-to-refresh or reward loops. • Compulsion loops: Repetitive gameplay cycles that exploit rewards and anticipation to keep players engaged. • Reduced Friction to Keep Playing/Paying: Design choices that minimize obstacles, encouraging seamlessness, continuous engagement and spending without reflection. • Artificial Scarcity: Time-limited Rewards • Escalating Rewards: Rewards increase in value or rarity the longer or more frequently a player engages, encouraging prolonged play and repeated investment. • Anticipatory Triggers: Cues or signals (like animations or sounds) build excitement before rewards, manipulating players' anticipation to boost engagement and spending. • Loot Box Rewards: Randomized Rewards [30]

			• Reward Saturation: Overexposure to frequent in-game rewards diminishes their perceived value, leading players to seek increasingly intense stimuli to stay engaged [31] • Friction in Account Deletion: Deliberate obstacles like hidden settings, multiple confirmations, or delays or making deleting accounts difficult, users discouraged from leaving a platform or game [32] • Bright Patterns Replacement: Substituting transparent, user-friendly defaults for dark-patterned UI elements such as visible opt-outs, clear labels, and honest nudges [33]
Forced Action	Forces users to complete specific actions to access features or proceed.	• Forced Continuity • Forced Registration • Privacy Zuckering • Friend Spam • Granting permissions and Exploiting Interaction flows	• Forced Continuity: Free trials that auto-renew without clear notice. • Forced Registration: Must sign up before browsing content. • Privacy Zuckering: Pushing users to share more data than necessary [34] • Friend Spam: Encouraging users to spam contacts or share invites [35] • Granting permissions and Exploiting Interaction flows: Requiring excessive permission to use a service.
Interface Interference	Manipulates visual hierarchy, wording, or layout to mislead users.	• False Hierarchy • Bad Defaults • Visual Prominence • Small or Moving Close Button • Trick Questions	• False Hierarchy: Highlights or enlarges options like “Accept All”. • Bad Defaults: Pre-selected options benefit the service provider. • Visual Prominence: Important choices are visually downplayed. [36] • Small or Moving Close Button: Makes it hard to close pop ups or ads. • Trick Questions: Confusing language in opt-in/opt-out options. [37]

Source: Authors, (2026).

III. APPLICATION DOMAINS OF DARK PATTERNS

As dark patterns have proliferated across digital ecosystems, research has increasingly focused on specific domains where manipulation is especially pervasive or impactful. Thematic analyses reveal that deceptive design strategies often evolve in tandem with platform-specific affordances, regulatory gaps, and commercial incentives. This section synthesizes findings across four major application areas: e-commerce and consent interfaces, mobile applications, social media platforms, and emerging modalities such as XR, voice interfaces, and games.

A. E-Commerce and Consent Interfaces

Dark patterns, specifically deceptive interface designs, are widely studied in e-commerce platforms due to their direct influence on consumer behavior and revenue generation. These platforms often deploy manipulative tactics such as hidden fees, urgency cues, misleading CTAs, and forced choices to increase conversions and secure consent while undermining transparency and user autonomy. Research on dark patterns highlights widespread manipulative design practices across digital platforms. Singh et. al. [38] audited Indian e-commerce platforms while Chaudhary et. al. [39] studied other E-commerce platforms and found urgency cues (“Only 1 item left!”), misdirection with prominent “Buy Now” buttons, and hidden fees at checkout, exploiting attention limits, creating artificial scarcity, and undermining transparency. Nouwens et al. [8], in a large-scale analysis of EU websites, identified pre-checked boxes, confusing language, and visually dominant “Accept All” buttons in cookie consent banners, showing that consent is manipulated rather than freely given.

Similarly, Santos et al. [40] examined web privacy interfaces and revealed deceptive layouts and text mimicking privacy tools while nudging consent, again demonstrating consent manipulation. Berens et al. [41] analyzed cookie disclaimers, identifying obscured opt-out paths and visual hierarchies favoring acceptance, which subvert informed consent and reduce transparency. Vishvakarma et al. [42] applied Agency Theory to e-commerce platforms, noting deception for profit, conflicts between business incentives and user interests, and information asymmetry that benefits platforms over users. Experimental evidence further illustrates these effects: Santana et al. [43] showed consumers often choose lower base-price options but end up paying more when hidden fees are added through drip or partitioned pricing and anchoring bias exploitation; although transparency reduces the effect, biases persist. Fecher [44] found that later disclosure of surcharges in travel-booking reduces perceived value and trust, particularly for price-conscious consumers, with drip pricing and hidden fees undermining brand trust and purchase intent. Moriuchi et al. [29] studied peer-to-peer lodging platforms like Airbnb,

showing that drip pricing, hidden fees, and manipulated fairness perceptions reduce consumer trust and intent when fees seem deceptive, though transparent pricing can lessen backlash. Finally, Ahmetoglu et al. [45] reviewed behavioral pricing tactics, finding that anchoring bias, partitioned pricing, drip pricing, and price obfuscation distort consumer choice, obscure total costs, reduce transparency, and raise ethical concerns about fairness in marketing. Together, these studies underscore how dark patterns exploit cognitive biases, manipulate consent, and erode consumer trust across contexts.

B. Mobile Apps and OTT Platforms

Mobile platforms introduce unique challenges for dark pattern detection and user experience due to small screen sizes, touch-based interaction, and app-centric ecosystems. These constraints are often exploited to hide options, obscure data deletion, or create misleading pathways. Dark patterns manifest differently across platforms and interfaces, often exploiting either consent, data, subscriptions, or interaction through manipulative navigation and behavioral nudges. Gunawan et al. analyzed cross-platform environments and identified bait and switch, hidden costs, forced continuity, and nagging, noting that limited mobile screen space intensifies hidden costs, preselection, and obfuscated opt-outs, disproportionately affecting users with cognitive or motor impairments. Ramteke et al. examined mobile apps and found friction in account deletion, misleading opt-out buttons, and covert data sharing practices; in freemium models, these deceptive designs drive retention while undermining informed consent and user control over data.

Similarly, Chen et al. used the AppRay tool to detect structural manipulation in app navigation, such as circular menus, disappearing back buttons, and trapped exit paths, showing how subtle UI designs hinder user autonomy and resist regulation. Ramokapane et al. [46] focused on data deletion workflows in mobile apps, uncovering multi-step deletion processes, non-intuitive instructions, and lack of confirmation or feedback, raising serious concerns about resistance to user data rights. In subscription contexts, Nguyen et al. [47] in his case study and Ye et al. [48] found that music streaming platforms like Spotify and Apple Music deploy auto-renewal defaults, friction in cancellation, and habit reinforcement loops, which promote lock-in, reduce autonomy, and influence behavior.

Oyibo et al. [49] investigated awareness of dark patterns in streaming sites such as YouTube and Ivysci, showing that users without knowledge are more susceptible to confirmshaming and trick questions (68% compliance), while awareness reduces compliance to 35%, indicating transparency and education mitigate manipulation and protect platform reputation. Nygren et al. [50] inspected interfaces across YouTube, Netflix, Disney+, and Prime Video, reporting widespread forced continuity, hidden subscription costs, invasive mobile pop-ups, manipulative cookie-consent defaults, and autoplay loops, underscoring the need for regulatory scrutiny and industry audits. Rhomberg et al. [51] and Alashwali et al. [52] emphasized interventions to empower users through tools like browser extensions, educational guides, and transparency features, proposing “bright patterns” such as auto-renewal toggles and clear consent mechanisms to reduce susceptibility.

Extending beyond traditional UI, Mattiuzzo et al. [53] showed that TikTok leverages personalization and design nudges algorithmic sequencing, timing controls, and personalized content loops to influence retention, sometimes overriding user intent, while Dubiel et al. [54] highlighted how enhanced voice fidelity in conversational agents increases trust and compliance even with misleading suggestions, making vocal manipulation a novel non-visual dark pattern that requires ethical regulation. Finally, Franzen et al. [55] explored UX practitioners’ perspectives, revealing that many rationalize privacy-related dark patterns such as default opt-ins, deceptive toggles, and coercive consent designs as necessary trade-offs, highlighting the urgent need for stronger ethical standards, designer training, accountability mechanisms, and regulation around privacy defaults.

C. Social Media and Engagement Loops

Social media platforms are designed around continuous engagement, and as such, they often employ dark patterns that target psychological vulnerabilities to maximize time-on-platform. These manipulations are deeply embedded in algorithmic curation, notification design, and interface dynamics. What distinguishes dark patterns in this domain is their temporal and psychological nature: rather than misleading a user once, they shape ongoing behavior, subtly steering attention and habit formation over time. This raises complex ethical issues, especially for younger users and vulnerable populations. Research on social media platforms shows how design patterns exploit psychological vulnerabilities to maximize engagement, often at the cost of user well-being and autonomy. Beltrán et al. [56] and Clark et al. [57] highlighted how infinite scroll, intermittent rewards such as variable likes, autoplay, and disappearing content create compulsive engagement loops, mirroring gambling mechanisms that exploit variable-ratio reinforcement.

These patterns reduce user well-being, induce anxiety, undermining control, and raise serious ethical concerns over loss of agency. Extending this critique, Mamun et al. [58] and Ahuja et al. [59] argued that platforms deliberately prioritize engagement optimization and manipulative metrics over user autonomy or mental health. They called for systemic reforms, including greater transparency in engagement metrics and regular audits of design practices, to ensure alignment with ethical standards. Together, these studies emphasize how social media dark patterns particularly those rooted in interaction design and reinforcement mechanics erode informed choice, compromise user agency, and necessitate stronger accountability mechanisms.

D. XR, Voice Interfaces, and Game Design

Emerging modalities like extended reality (XR), voice assistants, and digital games introduce novel interaction paradigms that also expand the scope of dark pattern deployment. In these environments, manipulation often exploits spatial, auditory, or reward-based structures that differ from traditional screens. voice-based dark patterns rely on tone, timing, and control of dialogue flow rather than visuals. They exploit listening effort, memory limits, and social cues to manipulate users’ decisions. The evolution of interfaces expands the scope of potential manipulative practices, extending beyond screen-based interactions to encompass bodily gestures, spoken commands, and reward-based systems. They also signal a growing need for modality-specific detection frameworks, legal protections, and design guidelines. Supportive vs. Manipulative Tones in Voice Interfaces. Research on online gaming highlights how dark patterns in design exploit psychological mechanisms, often undermining fairness, autonomy, and player well-being. Freeman et al. [60] examined

competitive games, showing that pay-to-win mechanics and purchase-driven power imbalances erode perceived fairness, reduce trust, and heighten frustration in competitive contexts.

Larche et al. [61] analyzed loot box systems, finding that rare rewards trigger heightened arousal and encourage repeated openings, exploiting variable reinforcement schedules akin to gambling and fostering potentially addictive behaviors. Vulnerable groups such as adolescents with Internet Gaming Disorder (IGD) face intensified risks. Liu et al. [62] revealed that reduced loss aversion and inhibitory control make them more susceptible to compulsion loops and low-friction play/pay mechanics, while Raiha et al. [63] demonstrated altered neural reward processing, where escalating rewards and anticipatory triggers reinforce addictive cycles. Similarly, Zhou et al. [64] found that IGD individuals display heightened sensitivity to in-game over real-world rewards, amplifying reward saturation and reinforcing escapism. Beyond addiction, engagement mechanics can also pressure players. Frommel et al. [65] studied daily quests and incentives, showing how artificial scarcity and fear of missing out (FOMO) can sustain habitual engagement but also generate burnout or pressure-driven play. Together, these studies illustrate how gaming dark patterns, especially through compulsion loops, reinforcement schedules, and reward manipulation, exploit cognitive vulnerabilities to drive monetization and engagement, raising ethical concerns over fairness, player health, and industry responsibility.

IV. AUTOMATED DETECTION OF DARK PATTERNS

As dark pattern research matured, scholars increasingly turned their attention to automated detection techniques aimed at identifying deceptive designs at scale. These efforts span mobile applications, websites, e-commerce systems, and emerging platforms, reflecting growing concern over unethical user interface practices. Approaches can be broadly categorized into language-based, visual/UI-based, and structural/DOM-based methods. Together, these methods reflect a maturing research landscape seeking scalable, generalizable, and explainable solutions.

A. Language-Based Detection

Language-based detection methods harness natural language processing (NLP) and transformer models to uncover manipulative textual content embedded in interfaces. Salminen et al. [66] trained a RoBERTa model on retail scenarios to identify deceptive strategies like drip pricing and false urgency. Their system achieved high F1-scores (up to 0.969), and the use of explainable AI tools such as LIME and SHAP enabled the identification of linguistic markers like scarcity, popularity, or FOMO cues. Sazid et al. [67] extended this line of work using GPT-3 in a few-shot learning setting, showing its ability to classify dark patterns with minimal labeled data, highlighting its adaptability across domains. Similarly, Ramteke et al. combined BERT with a web scraping pipeline to detect urgency-based and misdirection patterns in e-commerce, achieving a 96% accuracy rate. These studies underscore the effectiveness of transformer-based models in decoding subtle cues of manipulation in text-heavy interfaces.

B. Visual and UI-Based Detection

Because many dark patterns rely on visual deception—such as hiding options, contrast manipulation, or misleading layout—visual and UI-based approaches have emerged to interpret interface structure beyond language. Kodmurgi et al. [68] applied YOLOv5, a real-time object detection algorithm, to Android app screenshots, accurately identifying UI components associated with deceptive design choices. Mansur et al. [69] introduced AidUI, a multi-modal architecture that integrates visual layout features, user interaction traces, and semantic understanding to detect manipulation in both web and mobile contexts. These models highlight the importance of interpreting *how* interface components are positioned, styled, and presented to users—particularly when visual tricks are used to nudge behavior. Tools like UIGuard combine computer vision with text recognition and rule-based classification. Their system, built with Faster R-CNN and PaddleOCR, was deployed on mobile interfaces and achieved 0.93 accuracy. It significantly improved users' dark pattern detection recall (from 18.5% to 57.8%), though it remained constrained when handling unfamiliar pattern types. Likewise, Kodandaram et al. designed a detection system for screen-reader accessibility, targeting deceptive ads that harm visually impaired users. Their model combined transformer layers, an LSTM module, and DOM-derived features for high-precision classification, representing an often-overlooked domain in mainstream detection tools.

C. Structural/DOM-Based Detection

This approach focuses on analyzing a webpage's underlying logic and element hierarchy, rather than just its visible content or appearance. Early systems relied on rule-based parsing of HTML to flag patterns like "sneaking" or "misdirection." While these methods work well for known dark patterns, they require frequent manual updates and lack adaptability to new designs. To address these limitations, Nie et al. introduced the *Dark Pattern Analysis Framework* (DPAF), which builds on a taxonomy-driven review of 76 core studies. DPAF identifies 64 dark pattern types—20 of them newly defined—and maps existing detection methods to this taxonomy. Their work highlights significant interoperability issues and the absence of automated updating mechanisms that could keep pace with evolving deceptive design strategies. In a more holistic direction, Stavrakakis et al. introduced a morphological matrix framework that combines HTML parsing, computational linguistics, and image analysis. Their prototype successfully flagged patterns across categories but still faced challenges with UI. A third major category focuses on interface structure, particularly the Document Object Model (DOM) used in web and app design. These methods investigate how manipulation is embedded in the hierarchy, logic, or flow of interface elements. Bajaj et al. [70] modeled DOM trees in mobile apps to detect coercive flows—such as hidden cancel buttons or misleading progress indicators—by analyzing navigation paths and element nesting. Mansur et al. addressed the challenge of cross-platform applicability, proposing a generalization framework to train models that can detect dark patterns across varying UI architectures and environments. Cookie banners, a frequent site of consent manipulation, have been extensively studied through Document Object Model (DOM) analysis. Vedhapriyavadhana et al. [71] introduced a multi-faceted system to detect dark patterns such as deceptive UI/UX practices in shopping websites using Bidirectional Encoder Representations from Transformers (BERT) and DOM.

It used a Chromium browser extension, a machine learning backend, and a companion website to identify manipulative design elements in real-time and highlights them for users. Hausner et al. [72] used Support Vector Machines (SVMs) to classify manipulative designs based on HTML structure, textual cues, and CSS styling, with a focus on contrast tricks (e.g., highlighting "Accept" while hiding "Reject"). Ramokapane et al. applied machine learning to cluster banner components, revealing difficulties with inconsistent layouts and labeling schemes that hamper automated classification. These findings suggest that deception often resides in the interval variability and inconsistent semantics, echoing the call for more flexible, resilient systems. Together, these detection approaches illustrate the field's shift toward scalable, data-driven solutions. While many tools reach high performance in controlled environments, a common limitation persists difficulty generalizing to novel, dynamic, or platform-specific dark patterns. Emerging strategies aim to bridge this gap using transfer learning, explainable AI, and cross-platform models, but success will depend on continuously updated datasets, taxonomies, and more user-centered tools that promote transparency and trust. Despite promising accuracy, real-world deployment of these systems remains limited.

D. Hybrid/Ensemble Approaches

This approach combines two or more of the above strategies to maximize coverage. For example, a system may use NLP for text cues and DOM parsing for structural clues. Kodmurgi et al. introduced *ScrapeAI*, a multi-modal architecture that combines UI screenshots, code structure, and behavioral logs to detect dark patterns with promising accuracy. Abraham et al. proposed three detection algorithms—rule-based, ML-based, and hybrid—which showed complementary strengths in various commercial settings. Its multi-modal nature improves detection accuracy but comes at high computational cost and deployment complexity. In contrast, Abraham et al. advocate a modular hybrid model that balances rule-based interpretability with machine learning generalization. Their approach is more adaptable but struggles with dynamic content. Together, these methods reveal a trade-off between specificity and scalability—a challenge that remains unresolved.

V. MITIGATION STRATEGIES FOR DARK PATTERNS

This area remains underdeveloped compared to advances in detection. While research has focused heavily on identifying manipulative designs, fewer efforts address how to actively counter or prevent them. User-facing interventions show some promise: Schäfer et al. [73] demonstrated that simple design adjustments such as overlays, clearer consent options, and standardized button hierarchies can reduce mis clicks and enhance agency. Similarly, Kronhardt et al. [74] introduced a gamified tool that trains users to recognize manipulative patterns, improving awareness and literacy, though such approaches remain experimental and largely absent from mainstream platforms. From a regulatory and industry perspective, proposals like Rana et al. [75] Legal UX Audits suggest that interface designs could undergo compliance reviews much like financial or security audits. This model shifts the focus from penalizing violations after the fact to proactively setting standards for acceptable design. Yet these ideas remain largely theoretical, with limited institutional adoption. Meanwhile, a major gap persists for developers and designers, who lack real-time ethical evaluation tools comparable to accessibility or performance checkers. Without such support, ethical UI design often depends on ad-hoc expertise or retrospective audits, which are inadequate in fast-moving development contexts. Overall, progress in intervention is fragmented. To move forward, research and practice must embed ethical toolkits into design workflows, establish regulatory sandboxes, build educational infrastructure for users and developers, and standardize protocols for UI evaluation. Without these measures, detection will continue to outpace deterrence, allowing manipulative practices to evolve unchecked.

VI. LEGAL AND ETHICAL ANALYSIS

The legal and ethical analysis of dark patterns shows both convergence and divergence across global and Indian contexts. Globally, scholars like Luguri et al. frame the issue around autonomy and coercion, with regulation focused on consent, cookie banners, and manipulative e-commerce practices. Reports such as the Norwegian Consumer Council's 2018 [76] study spurred EU and US inquiries, and while frameworks like the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) set high standards, enforcement remains inconsistent. Studies indicate that platforms often practice strategic compliance rather than genuine design reform, leaving many forms of soft manipulation beyond effective legal reach. In India, by contrast, the framing emphasizes socio-technical vulnerability, reflecting the realities of mobile-first users, low digital literacy, and systemic inequities. The Consumer Protection Act (2019) [77] laid a broad foundation, but the 2023 "*Guidelines for Prevention and Regulation of Dark Patterns*" [78] marked a decisive step. It explicitly classifies manipulative practices and aligns with the Digital Personal Data Protection Act, emphasizing meaningful consent. Enforcement, however, is challenged by weak technical capacity, low consumer awareness, and fragmented oversight [79]. Ethically, the global debates stress autonomy and transparency, Indian discourse prioritizes digital equity, inclusive design, and protection for vulnerable users.

This comparison underscores that while the global approach leans on universal principles of autonomy, India advances a contextual, equity-driven model that addresses both cognitive and structural vulnerabilities [80]. Hence, combating dark patterns demands regulation, ethical design standards, user education, and metrics valuing long-term user welfare over short-term profit [81]. Even minor or small deceptive acts on digital platforms can erode trust [82]. Dark Patterns manipulate users' behavior, reduce autonomy, and undermine enjoyment, highlighting the urgent need for ethical, transparent design on various platforms [83]. Users still limit sharing sensitive data. Firms gain via better targeting, stronger marketing, and greater lock-in, showing regulation can protect privacy *and* support business outcomes [84][85][86]. Dark patterns continue to undermine consumer autonomy across domains such as e-commerce, gaming, and streaming. Hung (2021) highlights how "nagging" patterns create persistent consumer injury, while Li (2022) examines evolving regulatory responses under the FTC and CPRA to manipulative cookie consent notices. Few research focuses on remedies: Yao and Wei [87] propose improved terms-of-service design to reduce user confusion, whereas Singh, Vishvakarma, and Kumar [88] use grey influence analysis to study consumer challenges, and later authors in [89] show how deceptive design erodes brand equity. Zac et al. [90] stress that

consumer vulnerability magnifies harm, aligning with Bhoot et al. [91], who identify patterns through user reactions. Experimental evidence by Koh et al. [92] demonstrates how four common e-commerce dark patterns drive unintended consumption, while Nygren et al. document similar tactics in video streaming services. Parallel work on music streaming by Li, Liu, and Chen [93] underscores that user satisfaction and retention may be compromised by manipulative interfaces. Collectively, these studies suggest that dark patterns are not only prevalent but context-specific, raising urgent questions about ethics, user welfare, and regulatory oversight [94][95].

VII. METHODOLOGY

Step 1: Dataset Creation

A labeled dataset was created to examine 39 categories of dark patterns across widely used platforms. The categories include e-commerce (e.g., Amazon, Flipkart, Meesho, Myntra, Nykaa, Shopsy etc.), finance (Paytm, GPay etc.), gaming (Scribble.io, Minecraft etc.), matrimonial/dating/chatting (Instagram, whatsapp, Bhartmatrimony, Jeevansathi etc.), OTT (Netflix, Zee5, Amazon prime etc.), blockchain/cloud (AWS, GCP, casper, Blockchain.com etc.) and educational (Grammarly, W3 school and various universities) platforms. In this study, the unit of analysis was defined as an observable UI instance such as a text element, widget, flow step, or micro-interaction that exhibited or was suspected of exhibiting manipulative behavior (e.g., drip pricing, urgency cues, scarcity claims, or obstruction). To ensure comprehensive coverage, we examined key user flows across platforms, including the landing or homepage, sign-up and onboarding processes, the cart-to-checkout-to-pricing journey, cookie consent dialogs, settings and privacy options, as well as unsubscribe, refund, and cancellation mechanisms. These flows were tested across multiple user states namely, logged-out visitors, newly created accounts, returning users, and incognito sessions—to capture potential variations in dark-pattern deployment. A hybrid approach was employed, combining manual inspection of user flows with AI-assisted verification. Microsoft Copilot (Edge Sidebar) and Perplexity were used to extract UI text, summarize prompts, and flag potential dark patterns. Every AI prompt and output was logged for traceability, but manual verification remained mandatory to mitigate the risk of AI hallucinations and ensure the accuracy of identified instances.

A systematic process is followed to capture data to ensure consistency and reliability. For session setup, a Chromium-based browser (stable release) in both normal and private modes is used with cache and cookies cleared between sessions. The testing environment was configured to a desktop viewport ranging from 1366×768 to 1920×1080, set to the India locale with English as the default language. Only test accounts were created and used, with no personal data or payment details stored. For evidence collection, each candidate dark pattern was documented using multiple forms of data: screenshots and screen recordings captured before, during, and after user interactions; DOM snapshots that included visible text and relevant HTML snippets (via Inspect/Accessibility Tree); and, when necessary, network traces to verify whether timers or stock counters were dynamically fetched from the server or simply hard-coded and resettable. All captures were timestamped in ISO format to help detect recycled or false countdowns. These artifacts were stored in a structured directory hierarchy, with each session folder containing screenshots, recordings, DOM dumps, and notes for clarity and reproducibility. For AI-assisted review, Microsoft Copilot (Edge sidebar) and Perplexity were used as secondary verification tools. They extracted page text, summarized visible prompts, and suggested possible dark-pattern categories. However, their outputs were never accepted without manual confirmation, and any disagreements or vague categorizations were explicitly flagged as requiring manual adjudication. Candidate instances were assessed using objective and rule-based criteria to minimize subjectivity. A few examples are listed below in Table 2.

Table 2: Examples - Dark Patterns and their Trigger Rule.

Category	Objective	Decision Rule (Pass if…)	Example Triggers
Drip Pricing	Hidden fees, preselected add-ons	Material price element appears only after checkout initiation	“Convenience fee” at payment, gift wrap auto-selected
Urgency / Countdown	Timers, “ends soon” messages	Countdown persists across sessions or resets on reload	“Offer ends today” but reappears next day
Scarcity / Low Stock	“Only X left”	Stock claim static/unrealistic across refresh	“Only 1 left” on many SKUs
Social Proof / FOMO	“N people bought in 24h”	Unverifiable claims, often paired with urgency	“7,421 purchased today”
Disguised Ads	Sponsored content styled as organic	Paid content lacks equal disclosure	“Recommended” tiles w/o “Ad” tag
Friction in Deletion	Multi-step/hidden deletion	Requires >3 steps or off-platform action	“Email support to delete”
Forced Action	Pre-checked boxes, blocked progress	Opt-out harder to find/use than opt-in	“Yes, send me offers” checked by default
Hidden Costs	Late-stage fees	Non-optional fee not disclosed upfront	“Platform fee” added at last step

Source: Authors, (2026).

To evaluate candidate instances systematically, a set of rule checks was applied. The Equal-Prominence Test was used for consent-related choices, requiring that opt-out options carry comparable visual weight to opt-in options (at least 70% in size and positioned at the same panel depth). Failure to meet this criterion was marked as obstruction or misdirection. The Reset/Replay Test was applied to urgency and scarcity claims, where timers and stock counters were reloaded in fresh sessions; if these elements reset or behaved implausibly, they were flagged as false urgency or scarcity. Finally, the Click-Count Heuristic was used to evaluate cancellation flows, with subscribing allowed within two clicks but cancellations or deletions requiring four or more clicks or redirecting users off-platform are considered friction-inducing dark patterns.

Step 2: Cleaning of Dataset and Assigning weights using Saaty's AHP (Analytic Hierarchy Process)

For this study, a systematic data cleaning process was applied to ensure reliability and consistency of the binary dataset. First, all collected artifacts (screenshots, DOM snapshots, network traces, and notes) were cross-checked to confirm whether a dark pattern instance was genuinely present. Any ambiguous cases flagged by AI tools were manually reviewed and either validated or discarded to avoid false positives caused by hallucinated outputs. Duplicate records arising from repeated flows (e.g., refreshing pages or testing across user states) were identified and removed to prevent inflation of frequencies. Inconsistent labels across similar instances were standardized according to the predefined taxonomy of 39 dark-pattern categories, ensuring uniform coding. Finally, the binary encoding was verified where 1 consistently denoted the presence of a dark pattern and 0 denoted its absence, producing a clean and structured dataset suitable for pairwise comparison and AHP analysis. To construct the pairwise comparison matrix, the binary dataset was first interpreted. Each row represented an observation i.e., websites, webapps, mobile apps etc., belonging to different categories. Each column in the dataset corresponded to a criterion, such as Hidden Information, Obfuscation, Drip Pricing, Urgency, or Scarcity etc. Each criterion's frequency was then calculated as the number of rows in which it appeared (i.e., where the value equaled 1).

Using these frequencies, a reciprocal pairwise comparison matrix was generated, where the relative importance of criterion i over criterion j was defined as the ratio $a_{ij} = {f_i}/{f_j}$ where f_i and f_j represents frequencies of criterion i and j in the dataset. All diagonal elements (a_{ii}) are equal to 1 and $a_{ji} = {1}/{a_{ij}}$. The Saaty's Analytic Hierarchy Process (AHP) [96] is applied where the matrix was column-normalized, and the priority vector was computed as the row-wise average of the normalized matrix, yielding the relative weights w_i of each dark pattern. Finally, the consistency of the matrix was evaluated using AHP's consistency metrics. The largest eigenvalue λ_{max} was found to be 39.0 (equal to the number of criteria), leading to a Consistency Index (CI) of 0.0 and a Consistency Ratio (CR) of 0.0, indicating perfect consistency. This outcome was expected because the pairwise ratios were derived directly from observed frequencies, making the matrix mathematically transitive and free of contradictions, with the resulting weights directly reflecting the observed distribution of dark patterns in the dataset.

Step 3: Calculation of Risk Score, Trust Score and Diversity score using Weighted Shannon's Entropy Method

Trust Score, Risk Score, and Diversity Score of dark patterns across various categories in dataset are calculated using weighted Shannon's Entropy [97]. The algorithm begins by measuring the uncertainty of each criterion using equation 1 and equation 2. For every feature, Shannon entropy is computed by considering the distribution of its values across all records. Lower entropy indicates that the criterion provides more useful information, so it is transformed into an information utility measure by taking one minus the entropy value using equation 3. Next, it is combined with the Saaty AHP weights w_i computed in step 2. Each criterion's original weight w_i is multiplied by its information utility d_i and then normalized so that the adjusted weights sum to one. Using these adjusted weights w_i', the Risk Score for each row is calculated as the weighted sum of all the dark pattern indicators present, while the Trust Score is defined as one minus the Risk Score. Finally, the Diversity Score reflects the variety of dark patterns influencing each row and is computed as the Shannon entropy across that row's distribution of criteria. This way, Risk Score captures the intensity of dark patterns, Trust Score represents resilience against them, and Diversity Score measures how varied those dark patterns are within a given record.

To compute Trust Score (*TrustScore*), Risk Factor (*RiskScore*), and Diversity Score (*DiversityScore*) with Shannon entropy and AHP weights, with the dataset having m =40 rows (records) and n =39 criteria, with $x_{ij} \epsilon \{0,1\}$ indicating presence of criterion i in row j, and w_i is the AHP weight for criterion i , with $\sum_i w_i = 1$ calculated in step 2. The Shannon entropy of criterion i (normalized to [0,1]] for a binary variable) is given as equation 1:

$$H_i = -\left(\frac{1}{\ln(2)}\right)(p_i \ln(p_i) + (1 - p_i)\ln(1 - p_i)) \quad (1)$$

where p_i is calculated using equation 2 as:

$$p_i = \left(\frac{1}{m}\right)\Sigma_{j=1}^{m} x_{ij} \quad (2)$$

A small ε is used to replace any 0 inside logs if needed. The entropy is converted to information utility using equation 3 as:

$$d_i = 1 - H_i \quad (3)$$

The algorithm adjusts calculated weights using equation 4 to favor more informative criteria:

$$w_i' = \frac{(w_i * d_i)}{\sum_{\{k=1\}}^{\{n\}} w_{k^*} d_k} \quad (4)$$

Then the Risk score for each row i is the adjusted weighted sum of active criteria and is calculated using equation 5.

$$RiskScore_{\ i} = \sum_{i=1,j=1}^{m,n} w_j' * x_{ij} \quad (5)$$

The Trust Score for row i is calculated as complement of Risk Score using equation 6.

$$TrustScore_i = 1 - RiskScore_i \quad (6)$$

The Diversity Score for row i, is calculated using equation 7

$$.q_{ij} = \frac{\{x_{ij}\}}{\left\{\sum_{\{i=1\}}^{n} x_{ij}\right\};} \quad if \ \sum_{\{i=1\}}^{n} x_{ij} > 0$$
$$DiversityScore_i = -\left(\frac{1}{\ln(n)}\right) \sum_{j=1}^{n} q_{ij} \ln(q_{ij}) \tag{7}$$

The next section presents the analysis of results obtained after using above formulas on the dataset.

VII. RESULT ANALYSIS

The values of Risk Score (*RiskScore)*, Trust Score (*TrustScore)* and Diversity Score (*DiversityScore)* for our dataset are calculated using Shannon Entropy (weighted) method which uses Saaty's AHP for weight assignment. The analysis shows that e-commerce websites have the highest level of manipulation, with a very high *RiskScore* (~0.90) and very low *TrustScore* (~0.09). Their *DiversityScore* (~0.94) also suggests that they use wide variety of dark patterns, making them the most deceptive category overall. Educational websites present a more balanced picture, with a moderate *RiskScore* (~0.48) and a relatively higher *TrustScore* (~0.52). Their *DiversityScore* (~0.57) indicates that while some manipulative tactics exist, they are less frequent and less diverse compared to e-commerce. Finance platforms also show moderate levels, with *RiskScore* (~0.49) and *TrustScore* (~0.51) being almost evenly balanced. However, their *DiversityScore* (~0.73) suggests that financial platforms employ a broader set of manipulation techniques than educational sites, even if the overall risk is not excessively high.

Chatting and matrimonial/dating apps display a concerning profile, with a high *RiskScore* (~0.73) and a relatively low *TrustScore* (~0.27). Their *DiversityScore* (~0.87) highlights that they employ a large variety of dark patterns, making them one of the riskiest categories alongside e-commerce websites. Games are perceived far more positively, with a low *RiskScore* (~0.22) and a high *TrustScore* (~0.78). Their *DiversityScore* (~0.46) shows that although some manipulative designs exist (like loot boxes or addictive loops), they are relatively limited in range and intensity compared to other sectors. Finally, Blockchain and cloud platforms emerge as the safest and most trusted category. They have the lowest *RiskScore* (~0.18) and the highest *TrustScore* (~0.82) among all categories. Their *DiversityScore* (~0.37) indicates that they use very few dark patterns overall, reinforcing their perception as the least manipulative sector. These results are visualized using a heatmap and bar chart for different categories of website.

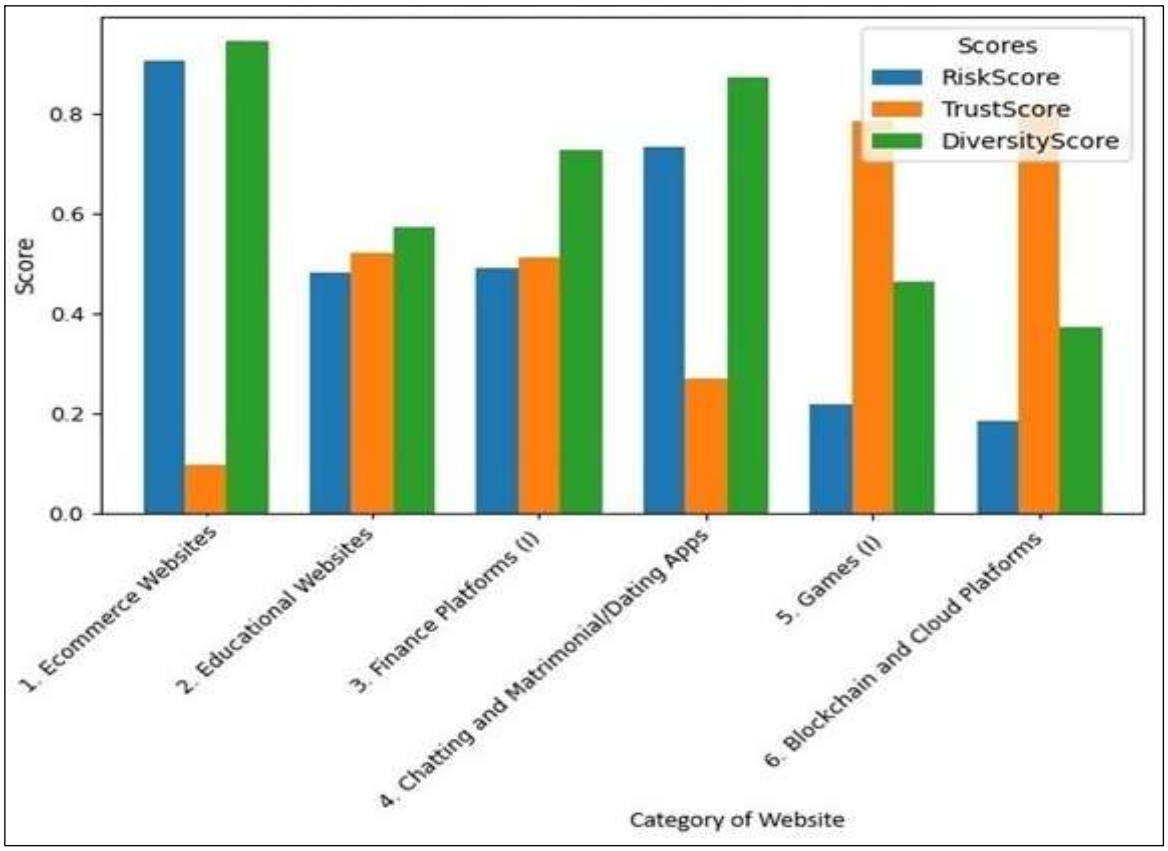


Figure 3:Category-wise Comparison Bar Chart.
Source: Authors, (2026).

The bar chart shows *RiskScore*, *TrustScore*, and *DiversityScore* for each category. In the heatmap, values are shown in matrix using color gradients instead of numbers. The heatmap highlights which scores are high/low with colors for quick comparison.

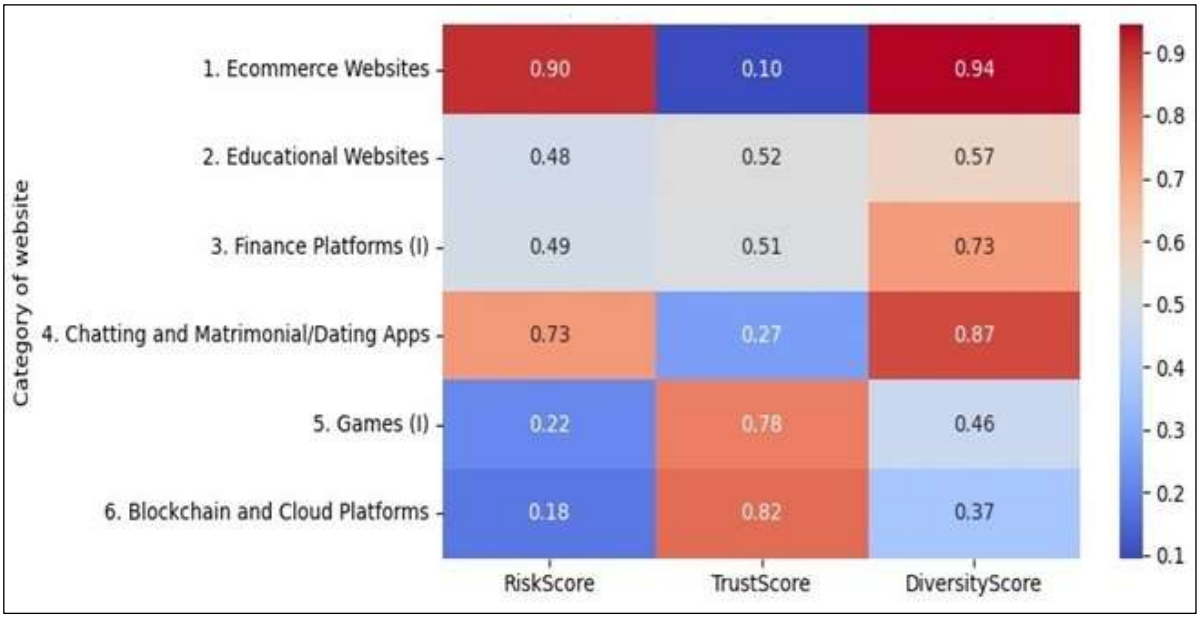


Figure 4: Category-wise Comparison Heatmap.
Source: Authors, (2026).

The charts in figure 3 and figure 4 reveal distinct sectoral differences in the prevalence and perception of dark patterns. E-commerce emerges as the highest-risk domain, characterized by the lowest levels of user trust and the widest variety of manipulative practices, reflecting how aggressive platforms optimize for sales. Dating apps also register as high risk with similarly low trust, as their designs often exploit emotional vulnerability and urgency. In contrast, finance and education platforms occupy a middle ground, showing moderate levels of risk, trust, and manipulation diversity due to stronger regulation in finance and ethical expectations in education. Interestingly, games and blockchain applications appear at the opposite end of the spectrum, with the lowest risk and highest trust, suggesting that users perceive these environments as more transparent or voluntary in nature, even though they operate in highly interactive and experimental domains. Overall, E-commerce and Dating Apps are the most manipulative, Finance and Education are moderately manipulative, while Games and Blockchain/Cloud Platforms are the least manipulative and most trusted.

V. CONCLUSIONS

To effectively mitigate dark patterns, a multi-stakeholder approach is essential. For designers and developers, the priority should be integrating ethical design toolkits and real-time pattern checkers into everyday workflows, alongside adopting standardized UI guidelines that reduce manipulative elements. Regulators and policymakers can play a crucial role by establishing regulatory sandboxes, where platforms can test interfaces for compliance, introducing legal UX audits or certification schemes modeled on GDPR and CCPA, and mandating enforceable design standards backed by penalties for non-compliance. Moreover, India's regulatory trajectory demonstrates a growing commitment to curbing digital exploitation, but sustained investment in technical expertise, consumer education, and coordinated enforcement is essential to ensure that these protections deliver on their promise. Users and communities should be empowered through digital literacy campaigns and gamified tools that enhance their ability to recognize manipulative interfaces, while participatory co-design with vulnerable groups, such as low-literacy, mobile-first, and multilingual population which ensure that regulations and design practices reflect diverse needs. Finally, researchers must shift focus from detection towards intervention, developing frameworks that can address dark patterns in emerging domains like extended reality, voice interfaces, and algorithmic platforms. Together, these strategies can move governance from reactive detection towards proactive prevention, fostering more transparent, equitable, and trustworthy digital environments.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Preeti Marwaha, Shalu Mahajan, Ishitva Joshi, Kriti Mishra, Ananya, Priyanka Sharma, Arunita Chaukiyal.
Methodology: Preeti Marwaha, Shalu Mahajan, Ishitva Joshi, Kriti Mishra, Ananya, Priyanka Sharma, Arunita Chaukiyal.
Investigation: Preeti Marwaha, Shalu Mahajan, Ishitva Joshi, Kriti Mishra, Ananya, Priyanka Sharma, Arunita Chaukiyal..
Discussion of results: Preeti Marwaha, Shalu Mahajan, Ishitva Joshi, Kriti Mishra, Ananya, Priyanka Sharma, Arunita Chaukiyal..
Writing – Original Draft: Preeti Marwaha, Shalu Mahajan, Ishitva Joshi, Kriti Mishra, Ananya, Priyanka Sharma, Arunita Chaukiyal.
Writing – Review and Editing: Preeti Marwaha, Shalu Mahajan, Ishitva Joshi, Kriti Mishra, Ananya, Priyanka Sharma, Arunita Chaukiyal.
Resources: Preeti Marwaha, Shalu Mahajan, Ishitva Joshi, Kriti Mishra, Ananya, Priyanka Sharma, Arunita Chaukiyal.
Supervision: Preeti Marwaha, Shalu Mahajan, Ishitva Joshi, Kriti Mishra, Ananya, Priyanka Sharma, Arunita Chaukiyal.
Approval of the final text: Preeti Marwaha, Shalu Mahajan, Ishitva Joshi, Kriti Mishra, Ananya, Priyanka Sharma, Arunita Chaukiyal.

VII. REFERENCES

[1] A. R. Casare, C. G. Da Silva, and R. Moraes, "User perception as a factor for improving trustworthiness in e-commerce systems," *Journal on Interactive Systems*, vol. 15, no. 1, pp. 194–219, 2024. doi: 10.5753/jis.2024.3748.

[2] J. Donia and J. A. Shaw, "Ethics and values in design: A structured review and theoretical critique," *Science and Engineering Ethics*, vol. 27, no. 5, p. 57, 2021.

[3] V. Bhaskaran, "Designing for trust: The crucial role in digital user experiences," *Journal of User Experience*, vol. 19, no. 2, pp. 53–59, 2024.

[4] L. Zhang-Kennedy, M. Keleher, and M. Valiquette, "Navigating the gray: Design practitioners' perceptions toward the implementation of privacy dark patterns," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–26, 2024. doi: 10.1145/3637374.

[5] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs, "The dark (patterns) side of UX design," in *Proc. CHI Conf. Human Factors in Computing Systems*, 2018, pp. 1–14.

[6] S. S. Chivukula and C. M. Gray, "Co-evolving towards evil design outcomes: Mapping problem and solution process moves," in *Proc. DRS*, 2020. doi: 10.21606/drs.2020.107.

[7] N. G. Devarmani *et al.*, "The fake discount epidemic in e-commerce platform: An examination of tactics, tools, and consumer awareness," *Cuestiones de Fisioterapia*, vol. 54, no. 1, pp. 796–807, 2025. doi: 10.48047/va8ybt47.

[8] M. Nouwens, I. Liccardi, M. Veale, D. Karger, and L. Kagal, "Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence," in *Proc. CHI Conf. Human Factors in Computing Systems*, 2020, pp. 1–13.

[9] D. Kelly and V. L. Rubin, "Identifying dark patterns in user account disabling interfaces: Content analysis results," *Social Media + Society*, vol. 10, no. 1, 2024, Art. no. 20563051231224269.

[10] R. F. Muhammad and S. Kasahara, "Agent-based simulation of fake news dissemination: The role of trust assessment and big five personality traits on news spreading," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 75, 2024.

[11] M. Botes, "Autonomy and the social dilemma of online manipulative behavior," *AI and Ethics*, vol. 3, no. 1, pp. 315–323, 2023.

[12] L. Alberts, U. Lyngs, and M. Van Kleek, "Computers as bad social actors: Dark patterns and anti-patterns in interfaces that act socially," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–25, 2024.

[13] A. Gerl and D. Pohl, "Critical analysis of LPL according to Articles 12–14 of the GDPR," in *Proc. 13th Int. Conf. Availability, Reliability and Security*, 2018, pp. 1–9. doi: 10.1145/3230833.3233267.

[14] Department of Consumer Affairs, *Draft Guidelines for Prevention and Regulation of Dark Patterns*. Ministry of Consumer Affairs, Government of India, 2023. [Online]. Available: https://consumeraffairs.nic.in

[15] J. Luguri and L. J. Strahilevitz, "Shining a light on dark patterns," *Journal of Legal Analysis*, vol. 13, no. 1, pp. 43–109, 2021. doi: 10.1093/jla/laz004.

[16] A. Mathur *et al.*, "Dark patterns at scale: Findings from a crawl of 11K shopping websites," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–32, 2019.

[17] A. Ramteke, S. Tembhurne, G. Sonawane, and R. N. Bhimanpallewar, "Detecting deceptive dark patterns in e-commerce platforms," *arXiv preprint* arXiv:2406.01608, 2024.

[18] J. Chen *et al.*, "Unveiling the tricks: Automated detection of dark patterns in mobile applications," in *Proc. 36th Annu. ACM Symp. User Interface Software and Technology*, 2023, pp. 1–20. doi: 10.1145/3586183.3606783.

[19] M. Hilton, "Dark patterns and user mental health: Identifying theoretical impacts of deceptive design on vulnerable demographics," in *Proc. Human Factors and Ergonomics Society Annu. Meeting*, vol. 67, no. 1, pp. 2124–2127, 2023.

[20] V. Krauß *et al.*, "What makes XR dark? Examining emerging dark patterns in augmented and virtual reality through expert co-design," *ACM Trans. Comput.-Hum. Interact.*, vol. 31, no. 3, pp. 1–39, 2024.

[21] J. Gunawan, A. Pradeep, D. Choffnes, W. Hartzog, and C. Wilson, "A comparative study of dark patterns across web and mobile modalities," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–29, 2021. doi: 10.1145/3479521.

[22] J. P. Zagal, S. Björk, and C. Lewis, "Dark patterns in the design of games," in *Foundations of Digital Games*, 2013.

[23] C. M. Gray, T. Mildner, and R. Gairola, "Getting trapped in Amazon's 'Iliad Flow': A foundation for the temporal analysis of dark patterns," in *Proc. CHI Conf. Human Factors in Computing Systems*, 2025, pp. 1–10. doi: 10.1145/3706598.3713828.

[24] D. Green, "Strategic indeterminacy and online privacy policies: (Un) informed consent and the General Data Protection Regulation," *International Journal for the Semiotics of Law*, vol. 38, no. 2, pp. 701–729, 2025.

[25] M. Degeling *et al.*, "We value your privacy... now take some cookies: Measuring the GDPR's impact on web privacy," *arXiv preprint* arXiv:1808.05096, 2018.

[26] J. Chen *et al.*, "From exploration to revelation: Detecting dark patterns in mobile apps," *arXiv preprint* arXiv:2411.18084, 2024.

[27] L. Nie *et al.*, "Shadows in the interface: A comprehensive study on dark patterns," *Proceedings of the ACM on Software Engineering*, vol. 1, no. FSE, p. 204–225, 2024. doi: 10.1145/364373.

[28] S. Santana, S. K. Dallas, and V. G. Morwitz, "Consumer reactions to drip pricing," *Marketing Science*, vol. 39, no. 1, pp. 188–210, 2020. doi: 10.1287/mksc.2019.1207.

[29] E. Moriuchi and S. Murdy, "Consumer reactions to drip pricing: The moderating effect of price fairness in the sharing economy accommodation," *Cornell Hospitality Quarterly*, vol. 66, no. 3, pp. 304–316, 2025. doi: 10.1177/19389655241271328.

[30] S. A. Goodstein, "When the cat's away: Techlash, loot boxes, and regulating 'dark patterns' in the video game industry's monetization strategies," *Univ. Colo. Law Rev.*, vol. 92, pp. 285–330, 2021.

[31] A. Singh, A. Arun, P. Malhotra, P. Desur, A. Jain, D. H. Chau, and P. Kumaraguru, "Erasing labor with labor: Dark patterns and lockstep behaviors on Google Play," in *Proc. 33rd ACM Conf. Hypertext and Social Media*, 2022, pp. 186–191. doi: 10.1145/3511095.3536368.

[32] B. Schaffner, N. A. Lingareddy, and M. Chetty, "Understanding account deletion and relevant dark patterns on social media," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–43, 2022.

[33] J. Aagaard, M. E. C. Knudsen, P. Bækgaard, and K. Doherty, "A game of dark patterns: Designing healthy, highly-engaging mobile games," in *Proc. CHI Conf. Human Factors in Computing Systems Extended Abstracts*, 2022, pp. 1–8.

[34] E. Dula, A. Rosero, and E. Phillips, "Identifying dark patterns in social robot behavior," in *Proc. 2023 Systems and Information Engineering Design Symp. (SIEDS)*, 2023, pp. 7–12.

[35] I. Stavrakakis, A. Curley, D. O'Sullivan, D. Gordon, and B. Tierney, "A framework of web-based dark patterns that can be detected manually or automatically," 2021.

[36] Y. Lu, C. Zhang, Y. Yang, Y. Yao, and T. J. Li, "From awareness to action: Exploring end-user empowerment interventions for dark patterns in UX," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–41, 2024. doi: 10.1145/3637336.

[37] W. Yi and Z. Li, "Mapping the scholarship of dark pattern regulation: A systematic review of concepts, regulatory paradigms, and solutions from an interdisciplinary perspective," *arXiv preprint* arXiv:2407.10340, 2024. doi: 10.48550/arXiv.2407.10340.

[38] V. Singh, N. K. Vishvakarma, H. Mal, and V. Kumar, "Prioritizing dark patterns in the e-commerce industry—an empirical investigation using analytic hierarchy process," *Measuring Business Excellence*, vol. 28, no. 2, pp. 177–192, 2024. doi: 10.1108/MBE-08-2023-0114.

[39] R. Chaudhary, S. Jain, R. Gupta, and V. Aggarwal, "Understanding the psychology of impulse buying in e-commerce: A behavioral review," *Journal of Marketing & Social Research*, vol. 2, pp. 102–113, 2025. doi: 10.61336/jmsr/25-06-13.

[40] C. Santos, N. Bielova, and C. Matte, "Are cookie banners indeed compliant with the law? Deciphering EU legal requirements on consent and technical means to verify compliance of cookie banners," *arXiv preprint* arXiv:1912.07144, 2019.

[41] B. M. Berens *et al.*, "Cookie disclaimers: Dark patterns and lack of transparency," *Computers & Security*, vol. 136, p. 103507, 2024. doi: 10.1016/j.cose.2023.103507.

[42] V. Singh, N. K. Vishvakarma, and V. Kumar, "Profit over principles: unveiling the motivating factors behind dark patterns in e-commerce through the lens of agency theory," *Journal of Enterprise Information Management*, vol. 38, no. 3, pp. 821–848, 2025.

[43] S. Santana, S. K. Dallas, and V. G. Morwitz, "Consumer reactions to drip pricing," *Marketing Science*, vol. 39, no. 1, pp. 188–210, 2020. doi: 10.1287/mksc.2019.1207.

[44] A. Fecher, T. Robbert, and S. Roth, "Same price, different perception: Measurement-unit effects on price-level perceptions and purchase intentions," *Journal of Retailing and Consumer Services*, vol. 49, pp. 129–142, 2019.

[45] G. Ahmetoglu, A. Furnham, and P. Fagan, "Pricing practices: A critical review of their effects on consumer perceptions and behaviour," *Journal of Retailing and Consumer Services*, vol. 21, no. 5, pp. 696–707, 2014. doi: 10.1016/j.jretconser.2014.04.013.

[46] K. M. Ramokapane, A. C. Mazeli, and A. Rashid, "Skip, skip, skip, accept!!!: A study on the usability of smartphone manufacturer provided default features and user privacy," *arXiv preprint* arXiv:2308.14593, 2023.

[47] F. Nguyen, "Trial length, pricing, and rationally inattentive customers," *arXiv preprint* arXiv:2507.06422, 2025.

[48] X. Ye, "Dark patterns and addictive designs," *Weizenbaum Journal of the Digital Society*, vol. 5, no. 3, 2025.

[49] K. Oyibo, "The influence of user knowledge and usage behaviour on decision-making and perceived reputation of streaming sites that use dark patterns," *Behaviour & Information Technology*, pp. 1–20, 2025. doi: 10.1080/0144929X.2024.2447475.

[50] F. Nygren and P. Tran, "Streaming in the dark: Analysing video streaming services for dark patterns: A user interface study," 2024.

[51] D. M. Rhomberg, "Dark patterns for good? Exploring end-user perspectives on bright patterns to counteract perceived social media problems," Ph.D. dissertation, Technische Universität Wien, 2024. doi: 10.34726/hss.2024.114637.

[52] E. Alashwali *et al.*, "Interface design to support informed choices when users face numerous privacy decisions," *IEEE Trans. Privacy*, 2025.

[53] M. Mattiuzzo and P. P. Ponce, "Power through design—dark patterns, personalization and the emergence of TikTok," *Int. Rev. Law, Comput. & Technol.*, vol. 39, no. 1, pp. 30–54, 2025. doi: 10.1080/13600869.2024.2351672.

[54] M. Dubiel, A. Sergeeva, and L. A. Leiva, "Impact of voice fidelity on decision making: A potential dark pattern?," in *Proc. 29th Int. Conf. Intelligent User Interfaces*, 2024, pp. 181–194. doi: 10.1145/3640543.3645202.

[55] D. Franzen, C. Müller-Birn, and O. Wegwarth, "Communicating the privacy-utility trade-off: Supporting informed data donation with privacy decision interfaces for differential privacy," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–56, 2024.

[56] M. Beltrán, "Defining, classifying and identifying addictive patterns in digital products," *IEEE Trans. Technol. & Soc.*, 2025.

[57] L. Clark and M. Zack, "Engineered highs: Reward variability and frequency as potential prerequisites of behavioural addiction," *Addictive Behaviors*, vol. 140, p. 107626, 2023.

[58] Q. Mamun, "Technology and social media's hidden cost: Social dilemma, mental health, misinformation, and manipulative practices," 2025.

[59] S. Ahuja and J. Kumar, "Conceptualizations of user autonomy within the normative evaluation of dark patterns," *Ethics and Information Technology*, vol. 24, no. 4, p. 52, 2022.

[60] G. Freeman, K. Wu, N. Nower, and D. Y. Wohn, "Pay to win or pay to cheat: How players of competitive online games perceive fairness of in-game purchases," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CHI PLAY, pp. 1–24, 2022. doi: 10.1145/3549510.

[61] C. J. Larche, K. Chini, C. Lee, M. J. Dixon, and M. Fernandes, "Rare loot box rewards trigger larger arousal and reward responses, and greater urge to open more loot boxes," *Journal of Gambling Studies*, vol. 37, no. 1, pp. 141–163, 2021.

[62] L. Wang *et al.*, "Reduced loss aversion and inhibitory control in adolescents with internet gaming disorder," *Psychology of Addictive Behaviors*, vol. 34, no. 3, pp. 484–494, 2020. doi: 10.1037/adb0000549.

[63] S. Raiha *et al.*, "Altered reward processing system in internet gaming disorder," *Frontiers in Psychiatry*, vol. 11, p. 599141, 2020. doi: 10.3389/fpsyt.2020.599141.

[64] W.-R. Zhou *et al.*, "Imbalanced sensitivities to primary and secondary rewards in internet gaming disorder," *Journal of Behavioral Addictions*, vol. 10, no. 4, pp. 990–1004, 2021. doi: 10.1556/2006.2021.00072.

[65] J. Frommel and R. L. Mandryk, "Daily quests or daily pests? The benefits and pitfalls of engagement rewards in games," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CHI PLAY, pp. 1–23, 2022. doi: 10.1145/3549489.

[66] J. Salminen, M. Mustak, S.-G. Jung, H. Makkonen, and B. J. Jansen, "Decoding deception in the online marketplace: Enhancing fake review detection with psycholinguistics and transformer models," *Journal of Marketing Analytics*, 2025.

[67] Y. Sazid and K. Sakib, "Prevalence and user perception of dark patterns: A case study on e-commerce websites of Bangladesh," in *Proc. ENASE*, 2024, pp. 238–249.

[68] P. S. Kodmurgi *et al.*, "ScrapeAI: A multi-modal approach to detect dark patterns," in *Proc. 15th Int. Conf. Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1–6. doi: 10.1109/ICCCNT61001.2024.10723319.

[69] S. M. H. Mansur, S. Salma, D. Awofisayo, and K. Moran, "Aidui: Toward automated recognition of dark patterns in user interfaces," in *Proc. IEEE/ACM 45th Int. Conf. Software Engineering (ICSE)*, 2023, pp. 1958–1970.

[70] A. Bajaj, K. Uppal, R. Razdan, Y. Tuteja, A. Bhardwaj, and A. Abraham, "A comprehensive analysis for dark pattern detection using structural, visual and textual information," *Int. J. Comput. Inf. Syst. Ind. Manage. Appl.*, vol. 17, p. 12, 2025. doi: 10.70917/ijcisim-2025-0002.

[71] R. Vedhapriyavadhana, P. Bharti, and S. Chidambaranathan, "Detecting dark patterns in shopping websites—A multi-faceted approach using BERT," *Enterprise Information Systems*, 2025, Art. no. 2457961.

[72] P. Hausner and M. Gertz, "Dark patterns in the interaction with cookie banners," *arXiv preprint* arXiv:2103.14956, 2021.

[73] R. Schäfer *et al.*, "Fighting malicious designs: Towards visual countermeasures against dark patterns," in *Proc. CHI Conf. Human Factors in Computing Systems*, 2024, pp. 1–13.

[74] K. Kronhardt, K. Rolfes, and J. Gerken, "Trickery: Exploring a serious game approach to raise awareness of deceptive patterns," in *Proc. Int. Conf. Mobile and Ubiquitous Multimedia*, 2024, pp. 133–147.

[75] A. Rana, "UX audit: A comprehensive review of methodologies and best practices for evaluating user experiences," *Int. J. Arts Archit. & Design*, vol. 2, no. 1, pp. 53–54, 2024.

[76] Norwegian Consumer Council, *Deceived by Design: How Tech Companies Use Dark Patterns to Discourage Us from Exercising Our Rights to Privacy*, 2018.

[77] M. Ghosh and K. Sarma, "Understanding consumer rights and responsibilities through consumer protection act 2019," *Int. J. Manage.*, vol. 11, no. 11, 2020.

[78] N. Gupta and A. George, "Digital personal data protection act, 2023: Charting the future of India's data regulation," in *Data Governance and the Digital Economy in Asia*, Routledge, 2025, pp. 34–53.

[79] S. Guha and S. Matilal, "Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021—A reassessment of the contours and limits," *NUJS J. Regul. Stud.*, vol. 8, p. 32, 2023.

[80] M. Guruswamy, "Justice K. S. Puttaswamy (Ret'd) and Anr v. Union of India and Ors.," *Am. J. Int. Law*, vol. 111, no. 4, pp. 994–1000, 2017.

[81] A. Narayanan, A. Mathur, M. Chetty, and M. Kshirsagar, "Dark patterns: Past, present, and future: The evolution of tricky user interfaces," *Queue*, vol. 18, no. 2, pp. 67–92, 2020.

[82] X. Zhan, Y. Xu, N. Abdi, J. Collenette, R. Abu-Salma, and S. Sarkadi, "Banal deception human-AI ecosystems: A study of people's perceptions of LLM-generated deceptive behaviour," *arXiv preprint* arXiv:2406.08386, 2024. doi: 10.48550/arXiv.2406.08386.

[83] R. P. Dahlan and M. Susanty, "Finding dark patterns in casual mobile games using heuristic evaluation," 2022. doi: 10.33322/petir.v15i2.1151.

[84] M. G. de Matos and I. Adjerid, "Consumer consent and firm targeting after GDPR: The case of a large telecom provider," *Management Science*, vol. 68, no. 5, pp. 3330–3378, 2022. doi: 10.1287/mnsc.2021.4054.

[85] C. F. Mondschein, "Some iconoclastic thoughts on the effectiveness of simplified notices and icons for informing individuals as proposed in Article 12 (1) and (7) GDPR," *Eur. Data Prot. L. Rev.*, vol. 2, p. 507, 2016.

[86] I. Gupta and P. Naithani, "Transparent communication under Article 12 of the GDPR: Advocating a standardised approach for universal understandability," *Journal of Data Protection & Privacy*, vol. 5, no. 2, pp. 149–161, 2022.

[87] J. Yao and H. Wei, “Creating better-informed consumers and reducing dark pattern tendencies through improved terms of service solutions,” 2018.

[88] V. Singh, N. K. Vishvakarma, and V. Kumar, “Investigating consumer challenges against dark patterns using grey influence analysis (GINA),” *Marketing Intelligence & Planning*, 2025. doi: 10.1108/mip-08-2024-0538.

[89] V. Singh, N. K. Vishvakarma, and V. Kumar, “Dark patterns, dimmed brands: The erosion of equity through deceptive design in e-commerce,” *Internet Research*, 2025. doi: 10.1108/intr-07-2024-1026.

[90] A. Zac, Y.-C. Huang, A. von Moltke, C. Decker, and A. Ezrachi, “Dark patterns and consumer vulnerability,” *Behavioural Public Policy*, pp. 1–50, 2023. doi: 10.1017/bpp.2024.49.

[91] A. M. Bhoot, M. A. Shinde, and W. P. Mishra, “Towards the identification of dark patterns: An analysis based on end-user reactions,” in *Proc. 11th Indian Conf. Human-Computer Interaction*, 2020, pp. 24–33. doi: 10.1145/3429290.3429293.

[92] W. C. Koh and Y. Z. Seah, “Unintended consumption: The effects of four e-commerce dark patterns,” *Cleaner and Responsible Consumption*, vol. 11, p. 100145, 2023. doi: 10.1016/j.clrc.2023.100145.

[93] Q. Li, Y. Liu, and C. Chen, “Satisfaction and continuation intention in music streaming services: Investigating key factors for user retention,” *Frontiers in Psychology*, vol. 16, p. 1552800, 2025.

[94] A. Hung, “Keeping consumers in the dark: Addressing ‘nagging’ concerns and injury,” *Colum. L. Rev.*, vol. 121, pp. 2483–2525, 2021.

[95] D. Li, “The FTC and the CPRA’s regulation of dark patterns in cookie consent notices,” *Univ. Chicago Business Law Rev.*, vol. 1, no. 1, pp. 19–42, 2022.

[96] R. W. Saaty, “The analytic hierarchy process—what it is and how it is used,” *Mathematical Modelling*, vol. 9, no. 3–5, pp. 161–176, 1987. doi: 10.1016/0270-0255(87)90473-8.

[97] C. E. Shannon, “A mathematical theory of communication,” *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001. doi: 10.1145/584091.584093.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 59-68. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2739

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

MODELING AND SIMULATION OF TEMPERATURE EFFECTS ON INGAP/INGAAS/GE TRIPLE-JUNCTION SOLAR CELLS WITH PYAMS

Fathi Dhiabi[*1] and Okba Benelmir[2]

[1]Laboratory of Metallic and Semiconducting Materials (LMSM), Biskra University, Biskra, Algeria.
[2]Department of Electronic and Automatic Engineering, Biskra University, Biskra, Algeria.

[1]https://orcid.org/0000-0001-7357-5946, [2]https://orcid.org/0009-0001-2167-6208

Email: *f.dhiabi@univ-biskra.dz, o.benelmir@univ-biskra.dz

ARTICLE INFO

Article History
Received: September 24, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
PyAMS,
Multi-junction solar cells,
InGaP/InGaAs/Ge,
Temperature effects,
Thermal modeling.

ABSTRACT

Multi-junction solar cells, particularly those based on *III–V* semiconductor materials; achieve high conversion efficiencies by employing multiple *P–N* junctions with different band gap energies to absorb distinct portions of the solar spectrum. Among them, the *InGaP/InGaAs/Ge* triple-junction photovoltaic cell stands out as a benchmark for efficiency. However, temperature variations strongly influence the electrical and thermal behavior of these subcells, often leading to performance degradation. In this work, we employ PyAMS to model and simulate the thermal response of InGaP/InGaAs/Ge solar subcells under varying temperature conditions. The developed model is validated against experimental data, ensuring accurate representation of the current–voltage (*I–V*) and power–voltage (*P–V*) characteristics. Key performance indicators—including short-circuit current density (Jsc), open-circuit voltage (*Voc*), fill factor (*FF*), and overall efficiency (η) are analyzed as functions of temperature. Our results demonstrate that while *Jsc* remains almost unaffected by temperature, *Voc* decreases significantly beyond 60°C under concentrated illumination (10 Suns). This drop in *Voc* strongly impacts both *FF* and *η*, highlighting the critical role of thermal management. The findings of this study provide useful guidelines for the design of cooling strategies in triple-junction solar cells, thereby improving their long-term performance, reliability, and integration into high-efficiency photovoltaic systems.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

III–V multi-junction solar cells are photovoltaic devices composed of multiple P–N junctions fabricated from different semiconductor materials with distinct band gap energies. Each junction absorbs and converts a specific portion of the solar spectrum into electrical energy, thereby enabling high conversion efficiency [1], [2]. The triple-junction photovoltaic (*TJ*) cell, consisting of Indium Gallium Phosphide (*InGaP*, top cell), Indium Gallium Arsenide (*InGaAs*, middle cell), and Germanium (*Ge*, bottom cell), is widely recognized as one of the most efficient architectures. However, like all electronic devices, the performance of these subcells is strongly influenced by temperature variations. Understanding their thermal behavior is thus crucial for optimizing performance and ensuring long-term reliability [3-6].

Realistic modeling of photovoltaic (*PV*) devices and accurate extraction of equivalent circuit parameters are of great importance for both design and operation, as they provide the foundation for future optimization strategies [7-14]. Thermal modeling, in particular, facilitates the development of efficient cooling systems for solar panels, helping to maintain high efficiency under concentrated illumination [15-18]. Among available simulation tools, PyAMS (Python for Analog and Mixed Signals) [19-22] has emerged as a powerful framework for studying the behavior of solar cells once their analytical equivalent circuits are established. PyAMS allows accurate representation of multi-junction devices while incorporating temperature dependence, making it possible to analyze critical performance parameters such as short-circuit current density (J_{sc}), open-circuit voltage (V_{oc}), fill factor (*FF*), and overall efficiency (η) [9].

This study focuses on investigating the performance of InGaP/InGaAs/Ge subcells under varying temperature conditions, with the objective of identifying the maximum allowable operating temperature that ensures optimal performance without degradation. To achieve this, the PyAMS software [22] is employed to develop a detailed model, which is then validated against experimental data. Subsequently, we analyze J_{sc}, V_{oc}, FF, and η as functions of temperature to determine the cooling requirements for stable operation.

The main contributions of this work can be summarized as follows:

1. Development of a PyAMS-based electrical and thermal model for InGaP/InGaAs/Ge triple-junction solar cells, incorporating temperature-dependent parameters such as saturation current, thermal voltage, and bandgap energy.
2. Validation against experimental data, confirming the accuracy of the model in reproducing *I–V* and *P–V* characteristics under different thermal and illumination conditions.
3. A comprehensive thermal performance analysis, showing that J_{sc} remains nearly stable with temperature, whereas V_{oc} decreases sharply above 60 °C under concentrated illumination.
4. Identification of the thermal operating limit of the Ge subcell (~60 °C at 10 Suns), which determines the overall system stability and directly affects fill factor and efficiency.
5. Providing practical design insights, offering guidelines for the integration of effective cooling strategies to enhance the long-term reliability of high-efficiency photovoltaic systems.

By addressing these aspects, this work advances beyond previous studies [3], [8], [9], [11], [13] by providing a validated PyAMS-based framework that simultaneously captures electrical and thermal behavior, while explicitly identifying practical cooling requirements for InGaP/InGaAs/Ge triple-junction solar cells.

II. SOLAR CELL MODELING

Modeling is a crucial step in understanding the electrical and thermal behavior of photovoltaic (*PV*) devices. In this section, we first discuss the equivalent circuit representation of a single-junction solar cell, which serves as the foundation for many studies in the literature. Then, we extend this representation into the PyAMS environment, highlighting its advantages in incorporating temperature-dependent relations and preparing the ground for triple-junction modeling [9], [23].

II.1 CLASSICAL EQUIVALENT CIRCUIT

As shown in Figure 1(a), the conventional single-diode equivalent circuit has been widely employed in the literature [9],[23],[24]. It consists of:

- A photocurrent source (I_{ph}), representing the current generated by photon absorption,
- A diode (*D*), describing the nonlinear behavior of the P–N junction according to the Shockley equation,
- A shunt resistance (R_{sh}), modeling leakage and parasitic currents across the junction, and
- A series resistance (R_s), accounting for ohmic losses in the semiconductor and electrical contacts.

This representation has proven effective in reproducing the current–voltage (*I–V*) behavior of solar cells. However, its classical formulation does not easily allow the integration of temperature-dependent effects, which are crucial for accurately modeling InGaP, InGaAs, and Ge subcells.

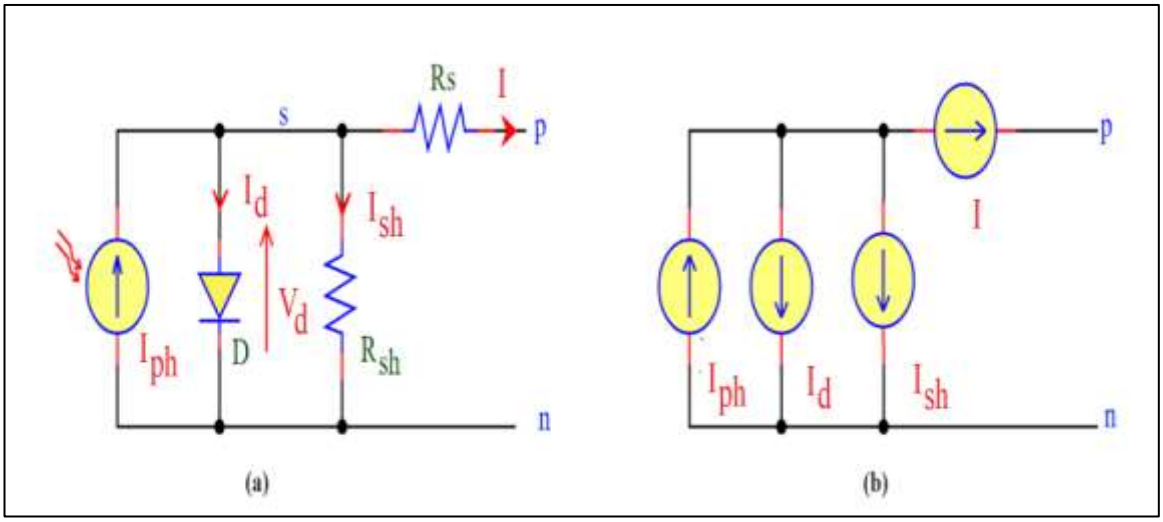


Figure 1: Equivalent electrical circuit of the solar cellM: (a) using discrete electronic components (b) using an independent current source as a function of voltage and parameters.
Source: Authors, (2026).

The photocurrent (I_{ph}) of the PV cell (Figure 1) depends on both the incident radiation and the temperature, as described by equation (1) [5], [23], [25], [26].

$$I_{ph} = [I_{sc} + K_i\,(T - 298)]\frac{Sun}{1000} \quad (1)$$

Where Isc represents the short-circuit current, K_i is a coefficient (*0.0017A/°C*), *T* denotes the temperature, and Sun stands for solar irradiance (W/m^2). The Id in Figure 1 is the current flowing through the diode D. The current is expressed by the following equation:

$$I_d = I_s\left(\exp\left(\frac{V_d}{V_t}\right) - 1\right) \tag{2}$$

For I_s [24], [27], [28], [29]:

$$I_s(T) = I_{so}\left(\frac{T}{300}\right)^3 exp\left(\left(\frac{T}{300} - 1\right)\frac{E_g}{V_t}\right) \tag{3}$$

The saturation current in the diodes of types InGaP, InGaAs, and Ge is expressed by the following formula [9], [23], [30]:

$$I_s(T) = I_{so}\left(\frac{T}{300}\right)^{\frac{3+\gamma}{2}} exp\left(\left(\frac{T}{300} - 1\right)\frac{E_g}{V_t}\right) \tag{4}$$

For V_t:

$$V_t = n\,\frac{KT}{q}. \tag{5}$$

In equations (2), (3), and (4), V_d represents the voltage applied across the diode, I_s is the reverse saturation current of the diode, q is the charge of an electron, V_t is the thermal voltage, K is Boltzmann's constant, T is the junction temperature, n is the ideality factor of the diode, I_{so} is the initial reverse saturation current of the diode, γ is a constant and E_g is the band gap energy of the semiconductor. The relationship between band gap energy and temperature is described by equation (6) below [8], [9], [10], [31]:

$$E_g(T) = E_g(0) - \frac{\alpha\, T^2}{T+\beta} \tag{6}$$

where α and β are a constant. The current in the resistor R_s or R_{sh} (Figure 1) is expressed by ohm's law:

$$I_{sh} = \frac{V_d}{R_{sh}} \tag{7}$$

and:

$$I = \frac{V_s}{R_s} \tag{8}$$

For V_s is the voltage applied across the resistor R_s. The open circuit voltage, V_{OC}, is the maximum voltage available from a solar cell, and this occurs at zero current. An equation for V_{OC} is found by setting the net current equal to zero in the solar cell equation to give [3]:

$$V_{oc} = \frac{nKT}{q}\ln\left(\frac{I_{sc}}{I_s} + 1\right) \tag{9}$$

II.2 PYAMS-BASED CURRENT-SOURCE MODEL

To overcome these limitations, we reformulate the equivalent circuit into a PyAMS-based current-source model, illustrated in Figure. 1(b). In this representation, each element of the solar cell is implemented as an independent current source defined as a function of voltage and physical parameters. This novel reformulation enables seamless integration of temperature-dependent relations into the model equations. The implementation in PyAMS is shown in Figure 2, where:

- The photocurrent (I_{ph}) is calculated as a function of irradiance, short-circuit current (I_{sc}), and temperature,
- The diode current (I_d) follows the Shockley relation with a temperature-dependent reverse saturation current $I_s(T)$,
- The shunt current (I_{sh}) is derived from the voltage across R_{sh}, and
- The series resistance contribution is expressed as an additional current component (I).

These currents are combined according to Kirchhoff's Current Law (*KCL*) within the `analog(self)` function in PyAMS. Temperature-dependent parameters such as the thermal voltage (V_t), the bandgap energy $E_g(T)$ (Varshni equation), and the reverse saturation current $I_s(T)$ are precomputed in the `start(self)` function before the simulation. Figure 3 presents the symbolic representation (symbol) of the single-junction solar cell in PyAMS. This symbol provides a reusable and modular block that can be directly inserted into larger circuit simulations, such as those involving triple-junction configurations.

```
# The single-junction solar cell solar cell model
class JunctionSolarCell(model):

    .......
    .......

  def start(self):
      # Thermal voltage at T (Eq. 3)
      self.Vt=vth(self.T)
      # Bandgap energy (eq.6)
      self.Eg=self.Ego-(self.α *self.T*self.T/(self. β +self.T))
      # Saturation current (Eq. 4)
      self.Is=self.Iso*pow(self.T, (self.γ +3)/2)*exp(-self.Eg/self.Vt)

  def analog(self):
      # Photocurrent equation (Eq. 1)
      self.Iph+= (self.Isc+self.Ki*(self.T-300))*self.Sun/self.Si
      # Diode equation (Eq. 2)
      self.Id+=self.Is*(explim(self.Vd/(self.n*self.Vt))-1)
      # Shunt resistor equation-high hom (Rsh) (Eq. 7)
      self.Ish+=self.Vd/self.Rsh
      # Series resistor equation-low hom (Rs) (Eq. 8)
      self.I+=self.Vs/self.Rs
```

Figure 2: The PyAMS model of a single-junction solar cell.
Source: Authors, (2026).

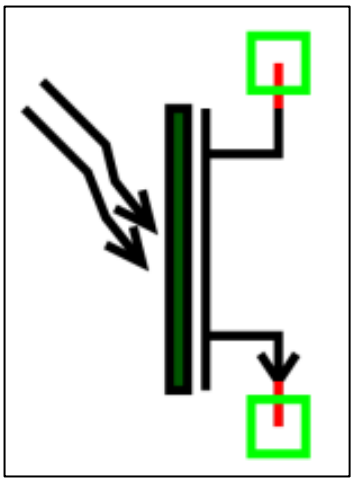


Figure 3: Single-junction solar cell symbol design by PyAMS.
Source: Authors, (2026).

II.3 CONTRIBUTION OF THE PYAMS REPRESENTATION

While Figure 1(a) reflects the standard baseline model used in most prior works, Figures 1(b), 2, and 3 illustrate the novelty of this study. By translating the conventional circuit into a PyAMS current-source framework, we enable:

- Direct integration of thermal effects, including $I_s(T)$ and $E_g(T)$,
- Modularity, as each current component is explicitly modeled and can be modified or extended, and
- Flexibility for multi-junction extension, since the single-junction block can be replicated and connected in series to form the InGaP/InGaAs/Ge triple-junction solar cell.

This approach bridges the gap between purely mathematical models and practical circuit-level simulations, offering a more accurate and flexible framework for analyzing both the electrical and thermal behaviors of high-efficiency photovoltaic devices.

II.4 MODELING OF TRIPL- JUNCTION SOLAR CELL

Building upon the single-junction framework, the structure of a triple-junction solar cell (TJSC) is illustrated in Figure 5(a). It consists of three subcells stacked in series:

- The top cell (InGaP) with a bandgap of approximately 1.9 eV,
- The middle cell (InGaAs) with a bandgap of 1.4 eV, and
- The bottom cell (Ge) with a bandgap of 0.6 eV [32-35].

This configuration is carefully designed so that each subcell absorbs a specific portion of the solar spectrum (Figure 4, right). The high bandgap top and middle subcells absorb shorter wavelengths, while the low bandgap Ge bottom cell absorbs longer wavelengths. This spectral division allows triple-junction devices to achieve significantly higher conversion efficiencies compared to single-junction solar cells [9]. The electrical model of the InGaP/InGaAs/Ge TJSC can be derived by connecting three single-junction submodels in series, each represented by its PyAMS-based equivalent circuit. This is shown in Figure 5(a), where the triple-junction equivalent circuit is constructed from three diode-based subcells. In the PyAMS implementation (Figure 5(b)), each subcell is replaced by its corresponding PyAMS model (developed in Section II.2). The series connection ensures that the same current flows through all three subcells, while the output voltage of the TJSC is the sum of the voltages of the individual cells. However, the lowest-performing subcell (typically the Ge bottom cell) determines the overall current, making it the limiting factor in device efficiency. By adopting this PyAMS-based modular approach, the TJSC model:

1. Accurately captures the spectral and electrical characteristics of each subcell,
2. Allows for the integration of temperature-dependent effects ($I_s(T)$, $E_g(T)$) for all three junctions, and
3. Provides a scalable framework that can be extended to simulate more advanced multi-junction architectures.

Thus, the novelty of our work lies in transforming the traditional equivalent circuit representation into a flexible PyAMS model, where each junction is explicitly defined and thermally coupled, enabling detailed analysis of the electrical–thermal behavior of the InGaP/InGaAs/Ge triple-junction solar cell.

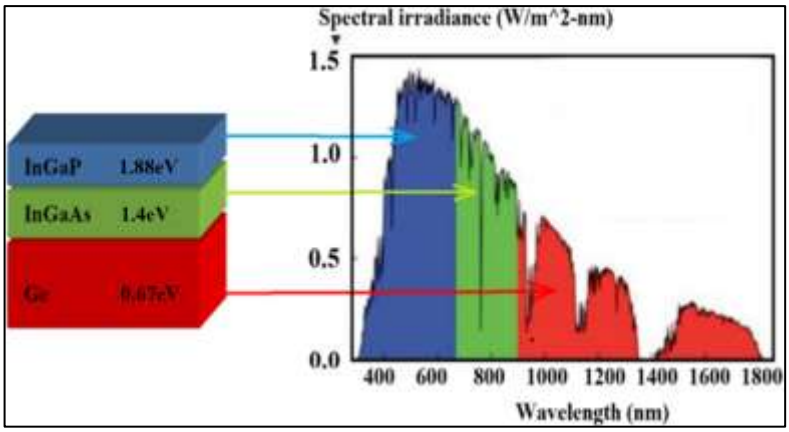
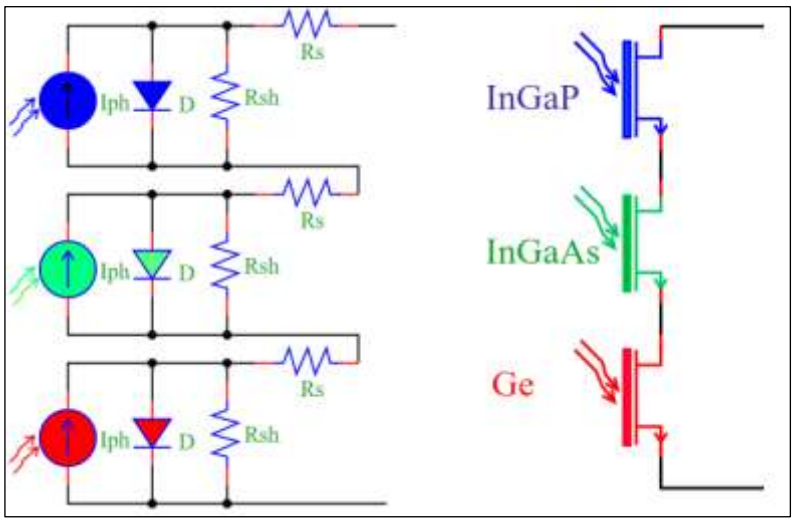


Figure 4: Triple junction solar cell: (a) Structure of a triple junction cell InGaP/InGaAs/Ge (b) The solar radiation spectrum versus wavelength and its useful spectrum portion.
Source: Authors, (2026).

This combination of materials was chosen because the P-N junction of each material would convert certain wavelengths of sunlight into electrical energy (Figure 5.b). Multi-junction solar cells with different band gap energies have high conversion efficiency [9]. The electrical modelling of the InGaP/InGaAs/Ge triple-junction photovoltaic cell can be implemented by the circuit shown in Figure 5(a) structure of circuit by triple circuit it presented in ref [8-11], where each sub-cell is represented by an independent single photovoltaic cell, in a way that gives the three cells the ability to be connected in series with decreasing gaps from top to bottom Figure 5(a) below shows the model for a triple junction solar cell, the PyAMS model of the triple-junction solar cell is developed based on this equivalent circuit, by replacing cell by equivalent Figure 5(b).



(a) **(b)**

Figure 5: The model of triple-junction Solar Cell: (a) Equivalent circuits (b) The PyAMS circuit.
Source: Authors, (2026).

III. SIMULATION RESULTS AND DISCUSSION

To validate the proposed PyAMS-based model, simulations were performed for the InGaP/InGaAs/Ge triple-junction solar cell and compared with published experimental data. The equivalent circuit of the three-junction device (Figure 5(b)) was implemented in PyAMS using the parameter values listed in Table 1, which were extracted under standard test conditions (air temperature of 27 °C and irradiance of 1000 W/m²).

Table 1: The parameters of the three junctions.

Parameter	Junction (InGaP)	Junction (InGaAs)	Junction (Ge)
Short circuit current I_{SC}	18mA	15.5mA	34mA
Bandgap energy E_g	1.89eV	1.40eV	0.683eV
Series resistance R_s	0.2190Ω	0.0219Ω	0.0219Ω
Shunt resistance R_{sh}	600kΩ	600kΩ	600kΩ
Alpha(α)	$7.50\ 10^{-4}$	$5.58\ 10^{-4}$	$4.77\ 10^{-4}$
Beta(β)	372	204	235
Gamma(Υ)	2	2	2
Ideality factor (n)	1.97	1.75	1.96
Saturation current I_o	$2.44\ 10^{-15}$	$3.39\ 10^{-11}$	$25\ 10^{-5}$

Source: Authors, (2026).

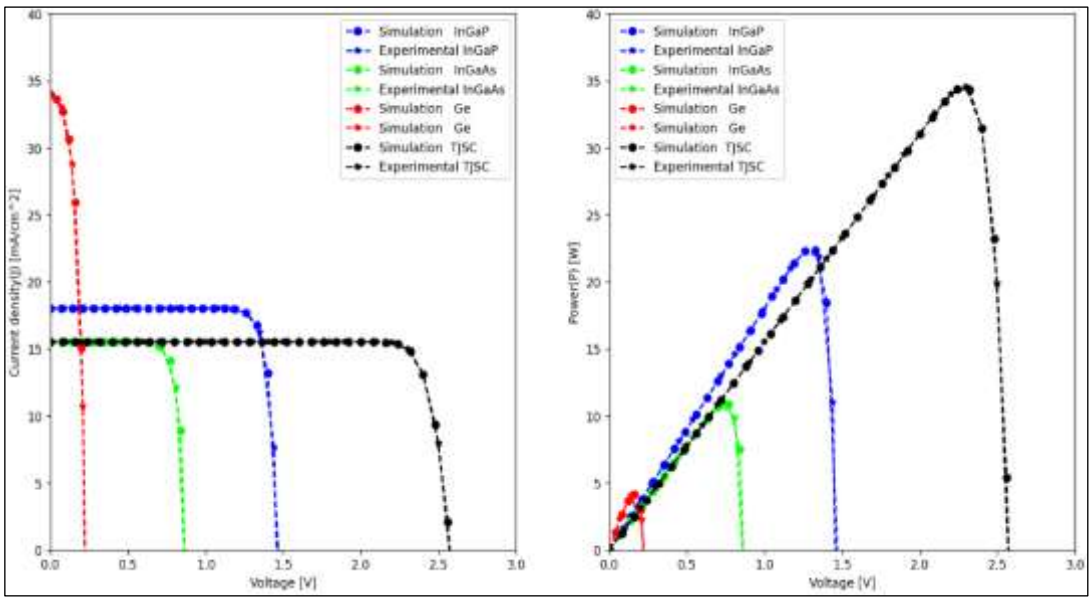


Figure 6: Validation of *I–V* (a) and *P–V* (b) characteristics of the individual bottom, middle and top subcells.
Source: Authors, (2026).

III.1 MODEL VALIDATION AGAINST EXPERIMENTAL DATA

The simulated current–voltage (I–V) and power–voltage (P–V) curves of the individual subcells (InGaP, InGaAs, Ge) and the full TJSC are shown in Figure 6. A close agreement is observed between simulated and experimental characteristics [32], confirming the accuracy of the PyAMS model. Quantitative performance metrics are summarized in Table 2, where the deviation between simulation and experiment is negligible (<1%). For example, the simulated short-circuit current density (Jsc) of the InGaP subcell (17.99 mA/cm²) is nearly identical to the experimental value (18.08 mA/cm²). Similar consistency is found for V_{oc}, *FF*, and η across all subcells and the complete *TJSC*.

Table 2: Performance parameters of simulated multijunction cell compared with experimental.

Junction	Result	Jsc(mA/cm²)	Voc(V)	Pmax(mW)	FF(%)	η (%)
InGaP	Simulation	17.99	1.50	23.18	85.37	23.18
	Experimental	18.08	1.49	22.9	84.90	23.00
InGaAs	Simulation	15.50	0.89	11.19	80.66	11.19
	Experimental	15.63	0.88	11.14	80.50	11.14
Ge	Simulation	33.99	0.24	4.59	54.31	4.59
	Experimental	34.23	0.25	4.57	54.17	4.57
TJSC	Simulation	15.76	2.60	34.50	84.15	34.90
	Experimental	15.79	2.58	34.33	84.06	34.33

Source: Authors, (2026).

This validation demonstrates that the PyAMS-based approach is capable of reproducing the performance of III–V multi-junction devices with high accuracy, comparable to experimental measurements.

III.2 TEMPERATURE-DEPENDENT ANALYSIS

To determine the maximum temperature tolerance of the sub-cells and the operating temperature efficiency of the triple-junction solar cell, we conducted simulations across a temperature range of 0°C to 80°C under concentrated illumination of 1000 Suns. The performance of each subcell analyzed as a function of temperature using PyAMS. The temperature limit deduced by observing the point at which the open circuit voltage dropped to almost zero.

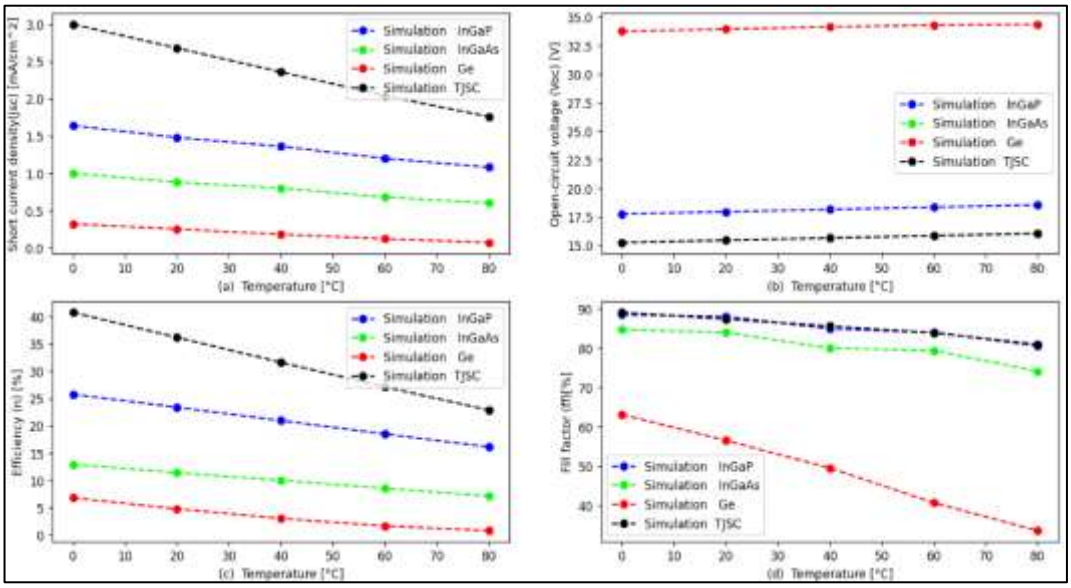


Figure 7: Effect of temperature on open-circuit voltage (a), current density (b), efficiency (c), and fill factor (d) for InGaP, InGaAs, Ge, and TJSC.
Source: Authors, (2026).

Figure 7(a) illustrates the short-circuit current density as a function of temperature of the three sub-cells and tandem cell. Interestingly, we observe no significant effect of temperature on the short-circuit current. In Figure 7(b), the open-circuit voltage plotted against temperature for the three sub-cells and the tandem cell. It is evident that the open-circuit voltage decreases significantly with increasing temperature. This decrease reflected in the slope, which diminishes across all concentration levels. Notably, for the Ge subcell, the open-circuit voltage approaches zero, resulting in a collapse of the current-voltage (*I-V*) curves for temperatures exceeding 60°C, whereas the upper and middle sub-cells can withstand temperatures exceeding 60°C. This decrease in open-circuit voltage significantly affects the fill factor and overall efficiency of the solar cell, as depicted in Fig. 7(c) and Figure 7(d). Consequently, the temperature tolerance range for this cell is limited to 0°C to 80°C. Figure 8 demonstrates this for the Ge subcell when exposed to various levels of solar radiation, showing that the open-circuit voltage completely disappears at a temperature of 60°C under concentrated illumination of 10 Suns.

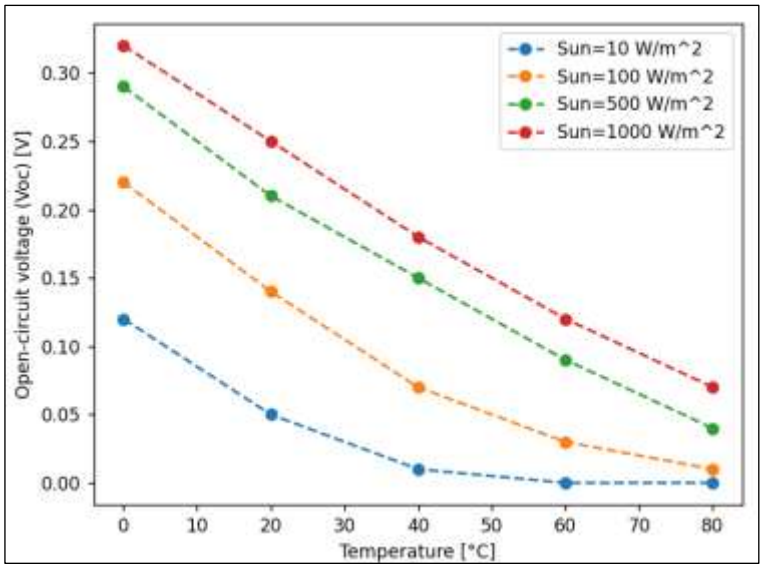


Figure 8: Effect of temperature and concentrated sunlight on open circuit voltage V_{oc} for bottom subcell.
Source: Authors, (2026).

After creating the InGaP/InGaAs/Ge equivalent model and comparing the experimental results in [32] with the same method of calculating the effect of temperature in [13], we obtained the same characteristics. Finally, we can conclude that the temperature limit is 60°C when exposed to concentrated illumination of 10 Suns, necessitating a cooling system to prevent the degradation of the InGaP/InGaAs/Ge MJSC.

III.3 DISCUSSION AND COMPARISON WITH PREVIOUS WORKS

The results highlight that while the upper InGaP and InGaAs subcells can tolerate elevated temperatures, the Ge subcell dictates the maximum operating limit of the TJSC. The identified threshold of ~60 °C under 10 Suns is consistent with experimental findings reported by Ota [9] and Djermane [13], but our model extends these works by providing a validated **PyAMS framework** that integrates both electrical and thermal effects. Unlike conventional SPICE-based or purely mathematical models, the PyAMS implementation offers two key advantages:

1. Explicit thermal coupling, by incorporating Is(T), Vt, and Eg(T) directly into the circuit-level model.
2. Modularity and scalability, enabling straightforward extension to multi-junction configurations while maintaining consistency with experimental behavior.

III.4 KEY FINDINGS

From this analysis, the following points can be emphasized:

- The PyAMS model reproduces experimental I–V and P–V data with high accuracy.
- Temperature has negligible influence on Jsc but a strong negative impact on Voc.
- The Ge bottom cell defines the thermal operating limit (~60 °C), beyond which FF and η rapidly degrade.
- Effective thermal management is essential to maintain stable performance of InGaP/InGaAs/Ge TJSCs under concentrated illumination.

IV. CONCLUSION

In this study, we developed and validated a PyAMS-based model for InGaP/InGaAs/Ge triple-junction solar cells. By reformulating the classical equivalent circuit into a modular current-source representation, the model successfully integrates temperature-dependent parameters such as saturation current, thermal voltage, and bandgap energy. Validation against experimental data confirmed that the PyAMS framework reproduces the *I–V* and *P–V* characteristics of individual subcells and the complete TJSC with high accuracy.Through systematic thermal analysis, we demonstrated that:

- The short-circuit current density (J_{sc}) remains nearly unaffected by temperature variations.
- The open-circuit voltage (V_{oc}) decreases significantly with increasing temperature, with the Ge subcell approaching zero Voc at ~60 °C.
- This drop in Voc strongly impacts the fill factor (*FF*) and overall efficiency (η), establishing the Ge subcell as the limiting factor in high-temperature operation.

These findings emphasize the critical role of thermal management in maintaining the stability and performance of multi-junction solar cells under concentrated illumination. The proposed PyAMS modeling framework not only provides accurate reproduction of experimental behavior but also offers a scalable and flexible tool for extending the analysis to advanced photovoltaic architectures. Future work will focus on integrating spectral effects, dynamic cooling strategies, and exploring the application of the PyAMS framework to other multi-junction material systems, thereby enhancing its potential as a versatile platform for photovoltaic device design and optimization.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Fathi Dhiabi and Okba Benelmir.
Methodology: Fathi Dhiabi.
Investigation: Fathi Dhiabi and Okba Benelmir.
Discussion of results: Fathi Dhiabi and Okba Benelmir.
Writing – Original Draft: Fathi Dhiabi.
Writing – Review and Editing: Fathi Dhiabi.
Resources: Fathi Dhiabi.
Supervision: Fathi Dhiabi.
Approval of the final text: Fathi Dhiabi and Okba Benelmir.

VI. REFERENCES

[1] M. Yamaguchi, "III–V compound multi-junction solar cells: present and future," *Sol. Energy Mater. Sol. Cells*, vol. 75, no. 1-2, pp. 261-269, January 2003. doi: 10.1016/S0927-0248(02)00168-X

[2] W. Guter, J. Schöne, S.P. Philipps, M. Steiner, G. Siefer, A. Wekkeli, E. Welser, E. Oliva, A.W. Bett and F. Dimroth, "Current-matched triple-junction solar cell reaching 41.1% conversion efficiency under concentrated sunlight," *Appl. Phys. Lett.*, vol. 94, no. 22, June 2009. doi: 10.1063/1.3148341

[3] N. Kensuke, T. Tatsuya, A. Takaaki, K. Minoru, U. Yukiharu and F. Takashi, "Evaluation of temperature characteristics of high-efficiency InGaP/InGaAs/Ge triple-junction solar cells under concentration," *Sol. Energy Mater. Sol. Cells*, vol. 85, no. 3, pp. 429-436, January 2005. doi: 10.1016/j.solmat.2004.05.008

[4] G.S. Kinsey, P. Hebert, K.E. Barbour, D.D. Krut, H.L. Cotal and R.A. Sherif, "Concentrator multijunction solar cell characteristics under variable intensity and temperature," *Prog. Photovolt: Res. Appl.*, vol. 16, no. 6, pp. 503-508, May 2008. doi: 10.1002/pip.834

[5] K. Nishioka, T. Takamoto, T. Agui, M. Kaneiwa, Y. Uraoka, T. Fuyuki, "Annual output estimation of concentrator photovoltaic systems using high-efficiency InGaP/InGaAs/Ge triple-junction solar cells based on experimental solar cell's characteristics and field-test meteorological data," *Sol. Energy Mater. Sol. Cells*, vol. 90, no. 1, pp. 57-67, January 2006. doi: 10.1016/j.solmat.2005.01.011

[6] G. Segev, G. Mittelman, A. Kribus, "Equivalent circuit models for triple-junction concentrator solar cells," *Sol. Energy Mater. Sol. Cells*, vol. 98, pp. 57-65, March 2012. doi: 10.1016/j.solmat.2011.10.013

[7] N. Das, H. Wongsodihardjo, and S. Islam, "Modeling of multi-junction photovoltaic cell using MATLAB/Simulink to improve the conversion efficiency," *Renewable Energy*, vol. 74, pp. 917-924, February 2015. doi: 10.1016/j.renene.2014.09.017

[8] Y. Sakurada, Y. Ota, and K. Nishioka, "Simulation of temperature characteristics of InGaP/InGaAs/Ge triple-junction solar cell under concentrated light," *Jpn. J. Appl. Phys.*, vol. 50, no. 4, pp.13-17 April 2011. doi: 10.1143/JJAP.50.04DP13

[9] Y. Ota, Y. Sakurada, K. Nishioka, "Temperature characteristics analysis of InGaP/InGaAs/Ge triple-junction solar cell under concentrated light using spice diode model," in *Proc. 35th IEEE Photovoltaic Specialists Conference, IEEE,* Honolulu, 2010, pp. 002093-002096. doi: 10.1109/PVSC.2010.5616163

[10] M. Yamaguchi, K. Nishioka1, T. Takamoto, T. Agui, M. Kaneiwa, Y. Uraoka, and T. Fuyuki, "Evaluation of InGaP/InGaAs/Ge triple-junction solar cell under concentrated light by simulation program with integrated circuit emphasis," *Jpn. J. Appl. Phys.*, vol. 43, no.3, pp. 882-890, March 2004. doi: 10.1143/JJAP.43.882

[11] Z. Song, Z. Wang, H. Zhang, W. Wu, B. Dou, Z. Tian, and C. Hu, "Influence of cell temperature on theoretical properties of InGaP/ InGaAs/Ge triple-junction concentrated solar cells ," *Recent Pat. Mech. Eng.*, vol. 17, no 3, pp.159-170, January 2024. doi: 10.2174/0122127976281400231211113923

[12] D. Kumar, N. Kumar, and L. Kumar, "Impact of Operating Temperature and Solar Concentration on the Conversion Efficiency of InGaP/InGaAs/Ge Hybrid Triple-Junction Solar Cell, " *Braz. J. Phys.*, vol. 53, no 1, pp. 31-41, December 2022. doi: 10.1007/s13538-022-01248-8

[13] K. Djermane, S. Kadri, A. Habbab, E. Bourbaba, Numerical Investigation of the Cooling Temperature of the InGaP/InGaAs/Ge Subcells Under the Concentrated Illumination," *Z. Naturforsch. A* vol. 75 no 2, pp. 93-101, November 2020. doi: 10.1515/zna-2019-0263

[14] X. You, Z. Wang,H. Zhang,B. Dou, W. Wu, and X. Zhou "Numerical simulation and experimental enalysis of electrical properties of InGaP/InGaAs/Ge triple-junction solar cell with reflective high-concentration photovoltaic system" [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3875188

[15] S. Anand, A. Gupta, S.K. Tyagi, Renew. Sustain. "Solar cooling systems for climate change mitigation: A review," *Energy Rev.*, vol. 41, pp. 143-161, January 2015. doi: 10.1016/j.rser.2014.08.042

[16] A. Alahmer, and S. Ajib, "Solar cooling technologies: State of art and perspectives," *Energy Convers. Manag.*, vol. 214, no. 15, June 2020. doi: 10.1016/j.enconman.2020.112896

[17] Y. Menni, A. Azzi, A.J. Chamkha, "A Review of Solar Energy Collectors: Models and Applications," *J. Appl. Comput. Mech.* vol. 4, no 4, pp. 375-401, June 2018. doi: 10.22055/jacm.2018.25686.1286

[18] K. Bataineh, Y. Taamneh, "Review and recent improvements of solar sorption cooling systems," *Energy Build.*, vol. 128, no. 15, pp. 22-37, September 2016. doi: 10.1016/j.enbuild.2016.06.075

[19] F. Dhiabi, M. Larbi Megherbi, A. Saadoune, R. Carotenuto, and F. Pezzimenti, "PyAMS: A new software for modeling analog elements and circuit simulations." *Int. J. Electr. Electron. Eng. Telecommun.* vol. 10, no. 4, pp. 233-242, July 2021. doi: 10.18178/ijeetc.10.4.233-242

[20] F. Dhiabi and M. Boumehraz, "Accelerating the solving of nonlinear equations using the homotopy method: Application on finding the operating point of complex circuits," *Turk. J. Elec. Eng. & Comp. Sci.*, vol. 25, no. 3, pp. 1864-1880, May 2017. doi: 10.3906/elk-1601-156

[21] F. Dhiabi. Design and Simulating Analog Circuits by PyAMS. [Online]. Available: https://zenodo.org/records/13823939.

[22] Python for analog and mixed signals (by using CAD system). [Online]. Available: https://pyams.sourceforge.io/

[23] A. Izadian, A. Pourtaherian, and S. Motahari, "Basic model and governing equation of solar cells used in power and control applications," in *Proc. IEEE Energy Convers. Congr. Expo. (ECCE), IEEE,* Raleigh, 2012, pp. 1483-1488. doi: 10.1109/ECCE.2012.6342639

[24] Y. Chaibi, A. Allouhi, M. Malvoni, M. Salhi, R. Saadani, "Solar irradiance and temperature influence on the photovoltaic cell equivalent-circuit models," *Sol. Energy vol.*, vol. 188, pp. 1102-1110, August 2019. doi: 10.1016/j.solener.2019.07.005

[25] S. Lineykin, M. Averbukh, and A. Kuperman, "Issues in Modeling Amorphous Silicon Photovoltaic Modules by Single-Diode Equivalent Circuit," *IEEE Trans. Ind. Electron.*, vol. 61, no 12, pp. 6785-6793, December 2014. doi: 10.1109/TIE.2014.2317138

[26] T.A. Chandel, M.Y. Yasin, and M.A. Mallick, "Modeling and Simulation of Photovoltaic Cell using Single Diode Solar Cell and Double Diode Solar Cell Model," *Int. J. Innovative Technol. Explor. Eng. (IJITEE)*, vol. 8, no. 10, pp. 558-565, August 2019. doi: 10.35940/ijitee.J8863.0881019

[27] M.Ş. Kalay, B. Kılıç, and Ş. Sağlam, "Systematic review of the data acquisition and monitoring systems of photovoltaic panels and arrays," *Sol. Energy*, vol. 244, pp. 47-64, September 2022. doi: 10.1016/j.solener.2022.08.029

[28] Y. Chaibi, A. Allouhi, M. Salhi, "A simple iterative method to determine the electrical parameters of photovoltaic cell," *J. Clean. Prod.*, vol. 269, art. no. 122363, October 2020. doi: 10.1016/j.jclepro.2020.122363

[29] X.H. Nguyen, M.P. Nguyen, "Mathematical modeling of photovoltaic cell/ module/arrays with tags in Matlab/Simulink," *Environ. Syst. Res.*, vol. 4, no. 1 pp. 2-13 December 2015. doi: 10.1186/s40068-015-0047-9

[30] A.B. Hussain, A.S. Abdalla,A.S. Mukhtar,M. Elamin,R. Alammari, and A. Iqbal, "Modelling and simulation of single- and triple-junction solar cells using MATLAB/SIMULINK" *Int. J. Ambient Energy*, vol. 38, no. 6, pp.613-621, November 2017. doi: 10.1080/01430750.2016.1181567

[31] A. Moliton, and J.M. Nunzi, "How to model the behaviour of organic photovoltaic cells," *Polym. Int.*, 55 no 6, pp. 583-600, June 2006. doi: 10.1002/pi.2038

[32] S. Kotamraju, S. Kotamraju, M. Sukeerthi, S.E. Puthan-veettil, M. Sankaran, "Study of degradation in InGaP/InGaAs/Ge multi-junction solar cell characteristics due to irradiation-induced deep level traps using finite element analysis," *Sol. Energy*, vol. 178, pp. 215-221, January 2019. https://doi.org/10.1016/j.solener.2018.12.036

[33] R.R. King, D.C. Law, K.M. Edmondson, C.M. Fetzer, G.S. Kinsey, H. Yoon, R.A. Sherif, N.H. Karam, "40% efficient metamorphic GaInP/GaInAs/Ge multijunction solar cell," *Appl. Phys. Lett.*, vol. 90, no. 18, pp. 16-18, April 2007. doi: 10.1063/1.2734507

[34] G.S. Kinsey, P. Hebert, K.E. Barbour, D.D. Krut, H.L. Cotal, R.A. Sherif, "Concentrator multijunction solar cell characteristics under variable intensity and temperature," *Prog. Photovolt: Res. Appl.*, vol. 16, no 6, pp. 503-508, September 2008. doi: 10.1016/j.solmat.2014.06.018

[35] M.K. Anaty, M. Belkasmi, K. Bouziane, M. Aggour, and M. Ouahab, "Modeling and simulation of a C3MJ+ triple junction solar cell using Matlab/Simulink" in *Proc. Int. Renewable Sustain. Energy Conf. (IRSEC), IEEE*, Morocco, 2016, pp. 752-757. doi: 10.1109/IRSEC.2016.7984015

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 69-76. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2740

RESEARCH ARTICLE OPEN ACCESS

PREDICTING CARDIOVASCULAR DISEASES THROUGH INTEGRATED DEEP LEARNING FEATURES

Islam D. S. Aabdalla[1]*, D.Vasumathi[2]

[1]Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Hyderabad (JNTUH), Hyderabad, Telangana, India
[2]Professor, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad (JNTUH), Hyderabad, Telangana, India

[1]https://orcid.org/0009-0008-9839-4576, [2]https://orcid.org/0009-0001-3610-0519

Email: * engislam2021@gmail.com, Vasukumar_devara@jntuh.ac.in

ARTICLE INFO

Article History
Received: September 25, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Machine learning,
Deep learning,
ECG,
Phonocardiogram,
Heart disease.

ABSTRACT

This research investigates the effectiveness of various machine learning models for classifying multimodal heart signals, specifically using electrocardiogram (ECG) and phonocardiogram (PCG) data from the EPHNOGRAM and PhysioNet Cardiology 2016 datasets. The proposed architecture, particularly the hybrid (LSTM, BiLSTM) model, achieved an impressive accuracy of 92.3% on the PhysioNet dataset, surpassing existing techniques and demonstrating significant potential for clinical applications. This study evaluates several architectures, including CNN, RNN, CNN and hybrid(RNN, LSTM), BiLSTM, and hybrid(LSTM, BiLSTM). By leveraging the EPHNOGRAM dataset, which contains synchronised ECG and PCG signals, the model aims to enhance classification accuracy for various heart conditions, including normal heartbeats, arrhythmias, murmurs, and other cardiovascular diseases (CVDs). Experimental results indicate that the proposed model outperforms traditional deep learning methods, achieving superior classification accuracy. These findings underscore the promise of multimodal deep learning models in healthcare diagnostics, particularly in detecting cardiac diseases. Future research will focus on optimising hyperparameters, utilising larger datasets, and integrating additional modalities, such as cardiac imaging and demographic data, to improve cardiovascular disease classification further.

Copyright ©2025 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Heart disease remains one of the leading causes of mortality worldwide, emphasising the importance of early detection to improve outcomes. Traditional diagnostic methods typically rely on either electrocardiogram (ECG) or phonocardiogram (PCG) signals, both of which provide vital insights into heart health. However, using only one modality limits the scope of diagnosis. ECG captures electrical activity and is effective for detecting arrhythmias, while PCG focuses on heart sounds, aiding in the diagnosis of murmurs and valve issues. Relying solely on one of these approaches can overlook complementary information[1]. Recent advances in machine learning, particularly deep learning, have opened new opportunities for analysing complex medical data, such as heart signals[2].Single-modality approaches often fail to provide a comprehensive diagnostic perspective. As highlighted by[3], multimodal methods that combine ECG and PCG signals significantly enhance diagnostic accuracy by leveraging their complementary strengths. This multimodal approach not only provides richer data but also helps reduce false positives by cross-validating information from different sources. Challenges in Multimodal Analysis Despite their benefits, the simultaneous processing of ECG and PCG signals presents challenges. Traditional machine learning models struggle to fully exploit the complex nature of multimodal data, as they require simultaneous extraction of spatial features from ECG and temporal features from PCG. Synchronisation and alignment issues between the signals further complicate this process. Standard algorithms relying on manual feature extraction are often insufficient for capturing the intricate relationships within multimodal data[4]. To address these limitations, this study proposes a hybrid deep learning model to advance heart disease diagnostics.

The model combines Convolutional Neural Networks (CNNs) to extract spatial features from ECG signals with Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM), to capture temporal sequences from PCG signals. Research by[5], demonstrates that hybrid CNN-RNN models outperform traditional approaches in multimodal contexts. By applying this strategy to ECG and PCG data, the proposed model aims to deliver more accurate and reliable diagnoses, addressing the shortcomings of single-modality systems. Research Objectives The primary goal of this research is to evaluate the effectiveness of various machine learning models in classifying multimodal heart signals, specifically focusing on ECG and PCG data. The study seeks to enhance classification accuracy for a range of heart conditions, including normal heartbeats, arrhythmias, murmurs, and other cardiovascular diseases (CVDs), using advanced deep learning architectures.

The main contributions include:

- Proposed a deep learning framework that integrates multiple architectures (CNN, RNN, LSTM, BiLSTM, and hybrid(LSTM, BiLSTM)) to classify multimodal heart signals.
- Demonstrated the effectiveness of synchronised ECG and PCG signals in enhancing classification precision.
- Improved classification capabilities for a wide range of heart conditions.
- Highlighted potential areas for future research, including optimising hyperparameters, leveraging larger datasets, and incorporating additional modalities like cardiac imaging and demographic information.

The structure of this paper is organised as follows: Section 2 presents the related work. Section 3 outlines the methodologies and techniques employed. Section 4 presents the results and discusses, and finally, Section 5 summarises the conclusions drawn from the research.

II. RELATED WORK

Deep learning has transformed cardiac diagnostics using ECG and PCG signals, improving the prediction of heart conditions. CNNs capture spatial patterns, RNNs handle temporal dynamics, and hybrid models show strong potential. Multimodal approaches combining ECG and PCG enhance accuracy, though challenges like data synchronisation persist. Cheng, et al[6]. Analysed both electrocardiogram (ECG) and phonocardiogram (PCG) signals. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), were widely used for cardiac diagnostics. CNNs demonstrated strong performance in feature extraction from ECG data due to their ability to capture spatial dependencies, while RNNs proved effective in handling the temporal nature of both ECG and PCG signals. In one study, a deep CNN combined with BiLSTM achieved notable improvements in ECG classification, reporting up to 98.7% accuracy in arrhythmia detection. According to et al[7]. Focused on PCG analysis using deep learning, which was less explored at the time but had great potential. A multimodal framework integrating both ECG and PCG signals was proposed to enhance diagnostic accuracy, especially under noisy conditions. However, challenges such as data synchronisation and feature alignment between ECG and PCG modalities persisted.

IDS Aabdalla, et al[8]. Explored ECG classification as a key approach for diagnosing heart disease, with a growing focus on prediction using machine learning techniques. The study aimed to improve cardiovascular disease classification accuracy by applying machine learning models to data from the PhysioNet database. They developed multi-class classification models to distinguish between heart failure rhythm (HFR), normal heart rhythm (NHR), and arrhythmia (ARR) using 162 ECG signals. Techniques such as frequency-time domain analysis and wavelet scattering were used for feature extraction. Their SVM model achieved a training accuracy of 97.1% and a testing accuracy of 92%, offering an efficient, accurate tool for heart disease diagnosis and a useful foundation for future research. Qiu, Yue, et al[9]. Examined multimodal learning approaches that combined multiple data types such as ECG and PCG. This direction was promising for cardiac diagnostics, as it allowed models to leverage complementary information from various physiological signals. A multimodal deep learning system was developed to integrate ECG and PCG signals for early detection of cardiovascular diseases. Despite the advancements, issues like data alignment, signal noise, and computational complexity limited the practical application of these multimodal systems, particularly in real-time scenarios like wearable and home-based monitoring systems.

Rai, Hari Mohan[10]. Demonstrated that hybrid architectures combining CNNs with RNNs were highly successful in analysing time-series data like ECG and PCG signals. For instance, a hybrid CNN-LSTM model proposed for ECG-based arrhythmia detection achieved an accuracy of 99.5% in recognising five types of arrhythmias. Another hybrid model combining CNN with BiLSTM showed superior performance by capturing both spatial and temporal patterns in ECG data, improving generalisation, and reducing overfitting across datasets. Islam, Md Shofiqul, et al. [11].Introduced the HARDC model, which used a dilated CNN and a bidirectional recurrent neural network (BiGRU-BiLSTM) architecture. By incorporating local and global feature information, alongside an attention mechanism, the model achieved a classification accuracy of 99.60%, with high F1 scores, precision, and recall. This approach also demonstrated reduced computation time, making it an innovative and cost-effective strategy for ECG signal compression and recognition. Negin, et al[12]. Developed a lightweight deep learning approach for detecting eight different cardiac arrhythmias and normal rhythms, achieving a diagnostic accuracy of 98.24%.

They used resampling and baseline wander removal techniques to preprocess ECG signals. The model, combining CNN and LSTM layers, was applied to ECG data from PhysioNet databases, demonstrating higher accuracy compared to previous methods. Rajakumar[13]. Proposed a novel approach for predicting fetal arrhythmias from cardiotocographic (CTG) signals using an Ensemble Feature Extraction (EFE) model. This model tackled the variability in CTG signals through sophisticated feature extraction methods and classification using the XGBoost algorithm. The model achieved an accuracy of 96.34%, precision of 96.7%, and recall of 97.3%, providing valuable insights into improving signal classification accuracy. Baharoon[14]. Enhanced multimodal deep learning systems by combining fundus images with demographic data (age and gender). Their model, HyMNet, outperformed other multimodal frameworks, achieving an AUC of 0.791 for heart disease prediction by fusing features from multiple sources, showcasing the potential of combining imaging with other data types in cardiac diagnostics.

Zeeshan, et al[15]. Presented two multimodal fusion frameworks, MIF and MFF, for ECG heartbeat classification by converting ECG signals into images and using deep learning techniques for feature extraction. These approaches significantly improved the detection of arrhythmias and myocardial infarctions, outperforming traditional methods by integrating diverse representations.
Suliman, I. D., & Vasumathi, D[2]. Highlighted the critical role of the heart in circulating blood throughout the body and addressed the global impact of cardiac disease (CD) as a leading cause of death. Detecting and predicting heart diseases remains challenging for healthcare professionals, given common but nonspecific symptoms such as rapid heartbeats, chest pain, and breathing difficulties. Early detection of CD can help individuals take preventive actions before the disease progresses. The study aimed to apply and compare various classification models, including K-Nearest Neighbours (KNN), Decision Trees (DT), Naive Bayes (NB), Logistic Regression (LR), and Random Forest (RF), using five cardiac disease datasets. The models were implemented in Anaconda Jupyter Notebook, utilising its robust libraries for greater accuracy.

After ten iterations, Logistic Regression (LR) achieved the highest accuracy of 96.67% on the Cleveland dataset, outperforming existing algorithms. Hemaxi, et al[16]. Explored deep learning models like Gated Recurrent Unit (GRU), LSTM, and CNN for cardiovascular disease classification using ECG data. Their research, using the PTB-XL dataset, found that 1D ECG representations were more effective in capturing temporal patterns essential for classification compared to 2D image-based approaches, particularly for GRU models, which exhibited superior sensitivity and specificity. Azemi[17]. Addressed the challenge of classifying heart sound recordings (PCGs) as normal, abnormal, or uncertain. The PhysioNet Challenge 2016 facilitated the development of algorithms for PCG classification, and top models used techniques like CNNs, AdaBoost, and SVMs, with sensitivity reaching 0.94 and specificity 0.78, providing valuable data for future research in heart sound classification. Zhou, F., & Fang[18]. Applied a fusion of image representations from ECG signals for arrhythmia classification, achieving an accuracy of 99.6%. Their method used recurrence plots (RP), Gramian angular fields (GAF), and Markov transition fields (MTF) to improve recognition accuracy, demonstrating the efficacy of image-based models in ECG classification.

Mohamed, et al[19]. Reviewed lightweight deep learning approaches, such as LSTMs and wavelet transforms, to enhance ECG signal classification efficiency. Their study proposed a simplified multimodal deep fusion model that reduced computational complexity while maintaining high accuracy (98.80%), particularly beneficial for small datasets. Shao, Zhe[20].Presented an Image Fusion Model (IFM) for ECG heartbeat classification. This model converted heartbeats into three distinct image types—GAF, RP, and MTF—merging them into a single modality. Using AlexNet for feature extraction, the model achieved high accuracy, precision, and recall, surpassing earlier approaches on the PhysioNet MIT-BIH and PTB datasets. While significant advancements have been achieved in deep learning-based cardiac diagnostics, especially concerning ECG data, there remains a critical need for further research in PCG analysis. Our studies have focused on various types of feature extraction, utilising different datasets and exploring diverse deep learning approaches.

III. METHODOLOGY

The proposed framework combines a CNN for spatial feature extraction and an RNN (specifically LSTM/BiLSTM) for capturing temporal dependencies from ECG and PCG signals. This hybrid model processes the data in two stages: first, it extracts key morphological patterns with CNN layers and then learning the sequential relationships between heartbeats through RNN layers. This integrated approach is designed to improve heart condition classification by leveraging both spatial and temporal features effectively.

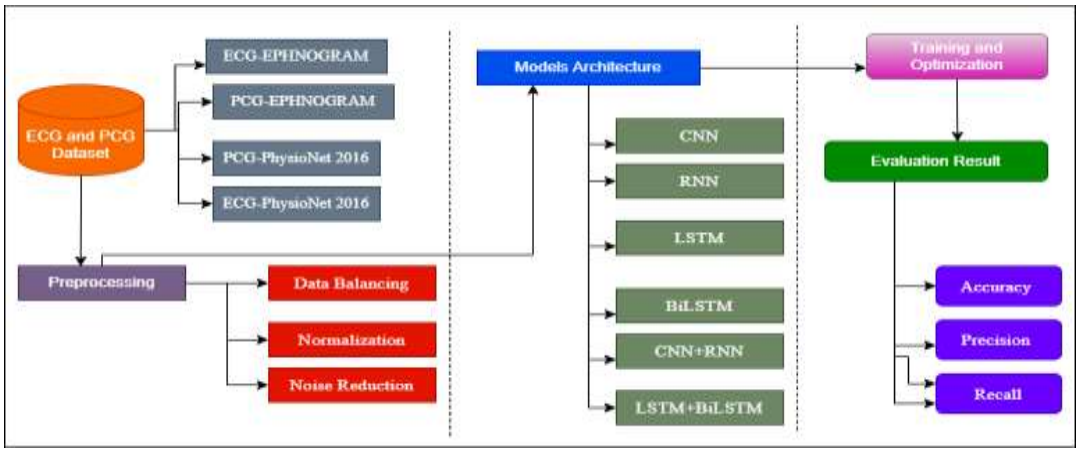


Figure 1: Overview of the Deep Learning Framework for ECG and PCG Analysis.
Source: Authors, (2026).

Dataset

The PhysioNet/Computing in Cardiology 2016 dataset contains over 3,100 recordings of ECG and PCG signals, which help in classifying heartbeats as normal, abnormal, uncertain, or artefact. These recordings span various durations and provide synchronised data crucial for multimodal analysis in heart disease detection[21]. The EPHNOGRAM dataset includes synchronised ECG and PCG signals from over 200 patients, capturing both the electrical activity of the heart (ECG) and its mechanical functions (PCG). It features a multimodal approach, which enhances classification performance for various cardiovascular conditions like arrhythmias and murmurs. The dataset is significant for its high sampling rate (typically 1,000 Hz) and the complementary nature of the data, making it highly valuable for detecting intricate heart conditions that may be missed using a single modality[21].

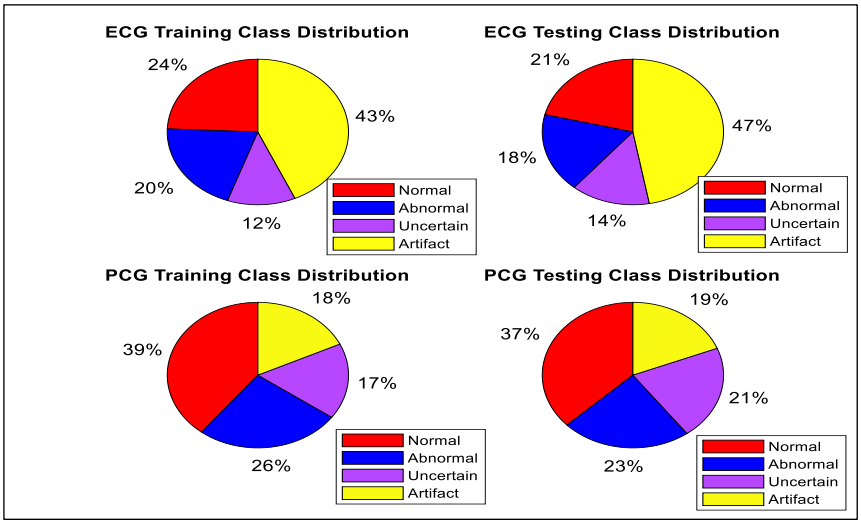


Figure 2: PhysioNet Cardiology 2016 Dataset.
Source: Authors, (2026).

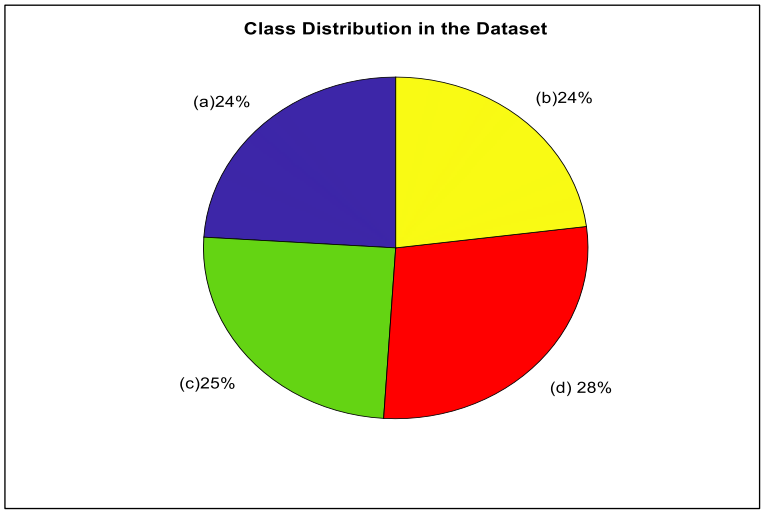


Figure 3: EPHNOGRAM Dataset.
Source: Authors, (2026).

Preprocessing

Data Balancing

Both the PhysioNet Challenge 2016 and EPHNOGRAM datasets exhibit class imbalance, with conditions like normal rhythms being more common than arrhythmias. This imbalance can lead to biased model predictions. To address this, oversampling techniques like SMOTE were used to enhance minority class representation, while undersampling was applied to reduce the size of the majority class. Data augmentation, including time-shifting, noise addition, and random cropping, was employed to increase the diversity of ECG and PCG signals while maintaining their physiological characteristics.

Signal Normalization

ECG and PCG signals from both datasets were collected from various devices, resulting in differences in amplitude and baseline drift. To standardise these signals, Min-Max normalisation scaled them to a [0, 1] range, while Z-score normalisation stabilised their distribution by adjusting for mean and standard deviation variations.

Noise Reduction

Noise in ECG and PCG signals, often due to muscle activity or stethoscope artefacts, was minimised using bandpass filtering. For ECG signals, frequencies between 0.5 Hz and 50 Hz were retained, while for PCG signals, a range of 20 Hz to 200 Hz was applied to preserve critical heart sound information and reduce unwanted noise.

Model Architecture

CNN for Feature Extraction:

CNNs are employed to extract spatial features from both ECG and PCG signals. The architecture consists of convolutional layers followed by ReLU activation and max-pooling layers, which downsample the data while preserving key features. This process reduces the input dimensionality and helps identify patterns in heart signals crucial for classification.

RNN for Temporal Learning:

After CNNs extract spatial features, RNNs, particularly Long Short-Term Memory (LSTM) or Bidirectional LSTM (BiLSTM), are used to capture the temporal relationships in the heart signals. These models are essential for processing sequential data like heartbeats, where dependencies between past and future information significantly influence the classification of heart conditions.

CNN+RNN Hybrid Model:

The hybrid CNN+RNN model leverages the spatial feature extraction capabilities of CNN and the temporal learning strength of RNN. This architecture efficiently captures both spatial patterns in the heart signals and the sequential dependencies over time, making it highly effective for analysing ECG and PCG data.

III.3 LSTM AND BILSTM MODELS:

LSTM captures long-term dependencies in ECG and PCG data by retaining key information from previous heartbeats, improving signal pattern analysis. BiLSTM extends this by processing data in both forward and backward directions, offering a deeper understanding of temporal dynamics. The LSTM+BiLSTM hybrid model combines both strengths, excelling at identifying complex patterns in ECG and PCG signals.

III.4 TRAINING AND EVALUATION

The training of the model was conducted using a cross-entropy loss function, which is commonly used for classification tasks. This loss function measures the performance of the classification model by comparing the predicted probability distribution to the actual class labels, penalising incorrect predictions.

Optimizer: The Adam optimiser was employed for training. Adam (Adaptive Moment Estimation) is popular because it combines the benefits of both the AdaGrad and RMSProp optimisers, using adaptive learning rates for each parameter to speed up convergence while maintaining stability.
Learning Rate: A fixed learning rate of 0.001 was used throughout training. This learning rate was chosen as a balance between training speed and model performance, allowing the model to converge without overshooting the optimal point.
The model parameters were updated iteratively based on the gradients calculated from the loss function, and training was conducted for multiple epochs to ensure that the model learned both spatial and temporal features effectively.
Evaluation Metrics:
The model's performance was evaluated using metrics, including:
Accuracy: The percentage of correctly predicted labels out of the total number of predictions. This metric indicates how well the model performs, but it may be less effective when classes are imbalanced.
Precision: The proportion of true positive predictions (correctly predicted positive class) out of all predicted positives. Precision is crucial for determining how well the model identifies true positives without mistakenly labeling negatives as positives.
Recall: The proportion of true positives out of the actual positives. It measures the model's ability to identify all relevant instances of the positive class.
F1-Score: The harmonic mean of precision and recall, providing a single metric that balances both. The F1-score is useful when the dataset has an imbalance between classes, since it penalises extreme values of either precision or recall.

IV.RESULT AND DISCUSSION

In our evaluation of the heart disease classification models using the PhysioNet Cardiology 2016 and EPHNOGRAM datasets, the hybrid LSTM+BiLSTM model achieved the highest performance, with an accuracy of 92.3%, precision of 91.9%, recall of 91.5%, and an F1-score of 91.7%. This suggests that combining long short-term memory networks with bidirectional LSTMs effectively captures temporal dependencies in the data. For the PhysioNet dataset, models based on CNN outperformed RNNs, highlighting the importance of convolutional feature extraction. In contrast, the performance of models on the EPHNOGRAM dataset was slightly lower overall, with the RNN+CNN model achieving a notable accuracy of 90.1%.

Table 3: Comparison of Model Performance DATASET 1.

DATASET	Models	Accuracy	Precision	Recall	F1-score
ECG and PCG PhysioNet Cardiology 2016	CNN	85.2	84.1	83.8	83.9
	RNN	88.5	87.9	88.1	88
	CNN	89.7	89.1	88.9	89
	RNN+CNN	90.4	90	89.8	89.9
	LSTM	91.1	91	90.7	90.8
	BiLSTM	92.3	91.9	91.5	91.7
ECG and PCG EPHNOGRAM	LSTM+BiLSTM	85.4	84.9	86.7	85.3
	RNN	88.2	87.5	89.1	88.3
	CNN	90.1	89.4	90.5	89.9
	RNN+CNN	84.7	83.5	85.2	84.3
	LSTM	86.9	86.2	87.1	86.6
	BiLSTM	89.4	88.8	89.9	89.3

Source: Authors, (2026).

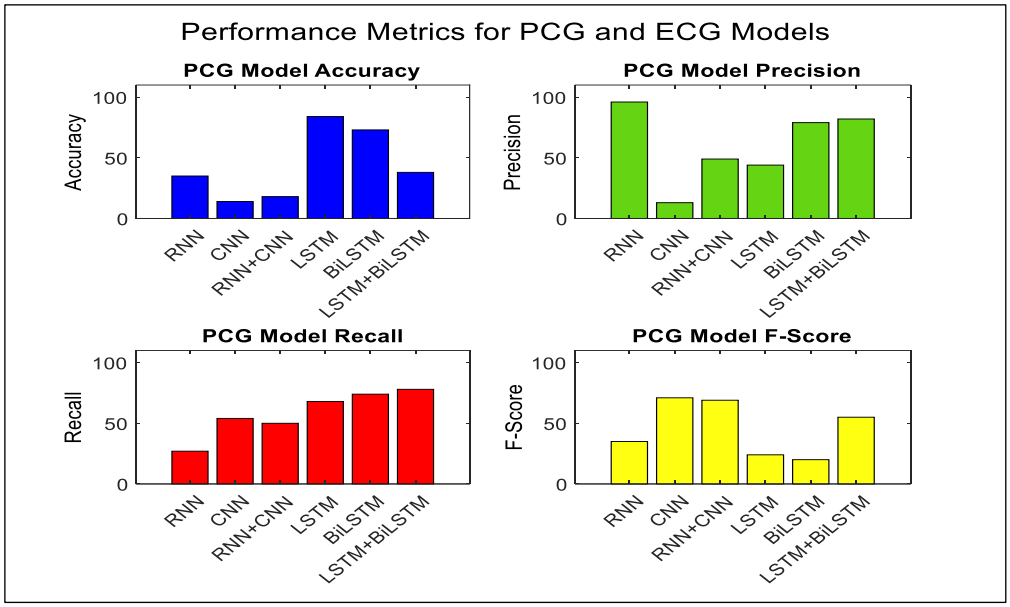


Figure 3: Performance Metrics of CVD Classification Models on the PhysioNet Cardiology 2016 Dataset.
Source: Authors, (2026).

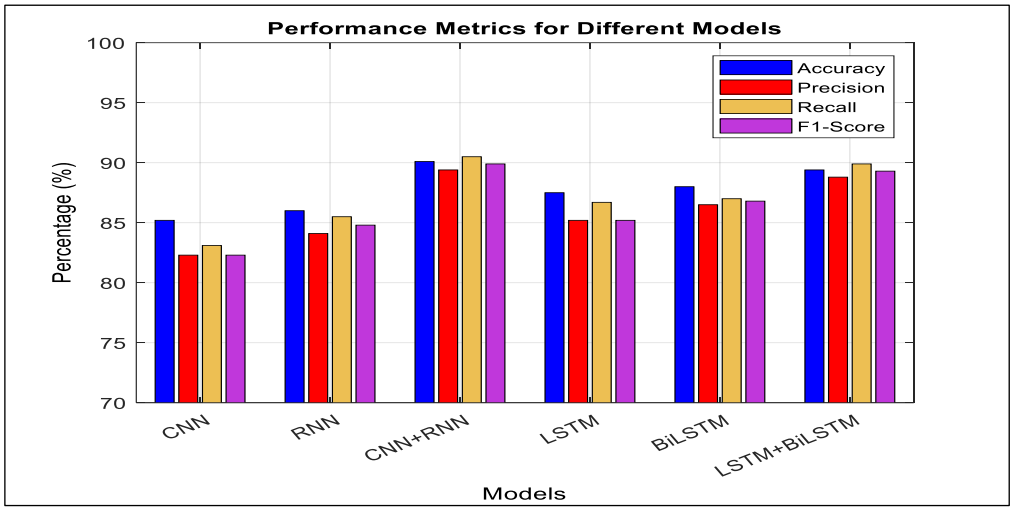


Figure 4: Performance Metrics of CVD Classification Models on the EPHNOGRAM Dataset
Source: Authors, (2026).

The table4 compares the performance of various models for classification tasks, highlighting accuracy, precision, recall, and F1-score. For instance, Diker et al.'s CNN achieved an accuracy of 83.82% with a precision of 82%, while Pengpai Li et al.'s LSTM+GA model excelled with an accuracy of 93.6%. Zeng et al.'s Bi-GRU model also performed well, achieving an F1 score of 93.27%. Our proposed LSTM+BiLSTM model achieved competitive results with an accuracy of 92.3%, a precision of 91.9%, and a recall of 91.5%. These results indicate its strength in handling time-series data, comparable to state-of-the-art methods like Zeng's Bi-GRU and Pengpai's LSTM+GA.

Table 4: Performance Comparison of Different Studies with Existing Work

Reference	Model	Accuracy	Precision	Recall	F1-score
A. Diker et al[1]	**CNN**	83.82	82	95	88.02
C. Pang et al[2]	**RF-RESNET**	87.2	88.8	79.5	83.9
Pengpai Li et al[3]	**LSTMs + GA**	93.6	90.3	84.5	87.4
Wajid Aziz et al[4]	**Ensemble Models**	81.9	89	83	85.8
Zeng, et al[5]	**Bi-GRU**	93.27	93.34	93.27	93.27
Ours	**LSTM+BiLSTM**	**92.3**	**91.9**	**91.5**	**91.7**

Source: Authors, (2026).

V.CONCLUSION

This research has demonstrated the effectiveness of various machine learning models for classifying multimodal heart signals, specifically using ECG and PCG data from the EPHNOGRAM and PhysioNet Cardiology 2016 datasets. The proposed architecture, particularly the LSTM+BiLSTM model, achieved a notable accuracy of 92.3% on the PhysioNet dataset, surpassing existing techniques

and highlighting its potential for clinical use. Future research will optimise hyperparameters, leverage larger datasets, and incorporate modalities such as cardiac imaging and demographics for the classification of cardiovascular disease.

VI.ACKNOWLEDGMENT

We sincerely express our gratitude to the Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Hyderabad (JNTUH), for providing the essential infrastructure, resources, and guidance that made this research possible.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Islam D. S. Aabdalla, D.Vasumathi.
Methodology: Islam D. S. Aabdalla, D.Vasumathi.
Investigation: Islam D. S. Aabdalla, D.Vasumathi.
Discussion of results: Islam D. S. Aabdalla, D.Vasumathi.
Writing – Original Draft: Islam D. S. Aabdalla, D.Vasumathi.
Writing – Review and Editing: Islam D. S. Aabdalla, D.Vasumathi.
Resources: Islam D. S. Aabdalla, D.Vasumathi.
Supervision: Mohammed Islam D. S. Aabdalla, D.Vasumathi.
Approval of the final text: Islam D. S. Aabdalla, D.Vasumathi.

VIII. REFERENCES

[1] H. Han, M. Xiang, C. Lian, D. Liu, and Z. Zeng, "A Multimodal Deep Neural Network for ECG and PCG Classification With Multimodal Fusion," in 2023 13th International Conference on Information Science and Technology (ICIST), 2023: IEEE, pp. 124-128.

[2] I. D. Suliman and D. Vasumathi, "Prediction of Heart Disease Using Machine Learning Algorithms," in 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2023: IEEE, pp. 1-12.

[3] K. Yan, T. Li, J. A. L. Marques, J. Gao, and S. J. Fong, "A review on multimodal machine learning in medical diagnostics," Math. Biosci. Eng, vol. 20, no. 5, pp. 8708-8726, 2023.

[4] N. Neelima, K. Jegdic, and A. Nagarwal, "Integrated Fusion Approach for Multi-Class HeartDisease Classification through ECG and PCG Signalswith Deep Hybrid Neural Networks," 2024.

[5] D. G. GONGADA, "MULTIMODAL SIGNAL ANALYSIS FOR DEPRESSION AND ANXIETY PREDICTION: A HYBRID CNN-RNN APPROACH," Journal of Theoretical and Applied Information Technology, vol. 101, no. 24, 2023.

[6] J. Cheng, Q. Zou, and Y. Zhao, "ECG signal classification based on deep CNN and BiLSTM," BMC medical informatics and decision making, vol. 21, pp. 1-12, 2021.

[7] S. A. L. Mary, S. Sivasubramanian, R. Palanisamy, and T. Thamizh Thentral, "An effective hybrid optimal deep learning approach using BI-LSTM and restricted Boltzmann machines whale optimization to detect arrhythmia," Multiscale and Multidisciplinary Modeling, Experiments and Design, pp. 1-19, 2024.

[8] I. D. Aabdalla and D. Vasumathi, "WAVELET SCATTERING TRANSFORM FOR ECG CARDIOVASCULAR DISEASE CLASSIFICATION."

[9] Y. Qiu et al., "Deep learning-based multimodal fusion of the surface ECG and clinical features in prediction of atrial fibrillation recurrence following catheter ablation," BMC Medical Informatics and Decision Making, vol. 24, no. 1, p. 225, 2024.

[10] H. M. Rai, K. Chatterjee, and C. Mukherjee, "Hybrid CNN-LSTM model for automatic prediction of cardiac arrhythmias from ECG big data," in 2020 IEEE 7th Uttar Pradesh section international conference on electrical, electronics and computer engineering (UPCON), 2020: IEEE, pp. 1-6.

[11] M. S. Islam, K. F. Hasan, S. Sultana, S. Uddin, J. M. Quinn, and M. A. Moni, "HARDC: A novel ECG-based heartbeat classification method to detect arrhythmia using hierarchical attention based dual structured RNN with dilated CNN," Neural Networks, vol. 162, pp. 271-287, 2023.

[12] N. Alamatsaz, L. Tabatabaei, M. Yazdchi, H. Payan, N. Alamatsaz, and F. Nasimi, "A lightweight hybrid CNN-LSTM explainable model for ECG-based arrhythmia detection," Biomedical Signal Processing and Control, vol. 90, p. 105884, 2024.

[13] S. Magesh and P. Rajakumar, "Ensemble feature extraction-based prediction of fetal arrhythmia using cardiotocographic signals," Measurement: Sensors, vol. 25, p. 100631, 2023.

[14] M. Baharoon et al., "HyMNet: a Multimodal Deep Learning System for Hypertension Classification using Fundus Photographs and Cardiometabolic Risk Factors," arXiv preprint arXiv:2310.01099, 2023.

[15] Z. Ahmad, A. Tabassum, L. Guan, and N. M. Khan, "ECG heartbeat classification using multimodal fusion," IEEE Access, vol. 9, pp. 100615-100626, 2021.

[16] H. Narotamo, M. Dias, R. Santos, A. V. Carreiro, H. Gamboa, and M. Silveira, "Deep learning for ECG classification: A comparative study of 1D and 2D representations and multimodal fusion approaches," Biomedical Signal Processing and Control, vol. 93, p. 106141, 2024.

[17] H. Hazeri, P. Zarjam, and G. Azemi, "Classification of normal/abnormal PCG recordings using a time–frequency approach," Analog Integrated Circuits and Signal Processing, vol. 109, no. 2, pp. 459-465, 2021.

[18] F. Zhou and D. Fang, "Multimodal ECG heartbeat classification method based on a convolutional neural network embedded with FCA," Scientific Reports, vol. 14, no. 1, p. 8804, 2024.

[19] M. Hammad, S. Meshoul, P. Dziwiński, P. Pławiak, and I. A. Elgendy, "Efficient lightweight multimodel deep fusion based on ECG for arrhythmia classification," Sensors, vol. 22, no. 23, p. 9347, 2022.

[20] Z. Shao, "Deep Atrial Fibrillation Classification Based on Multi-modal Attention Network," in Proceedings of the 2022 14th International Conference on Machine Learning and Computing, 2022, pp. 512-518.

[21] M. A. Reyna et al., "Heart murmur detection from phonocardiogram recordings: The george b. moody physionet challenge 2022," PLOS Digital Health, vol. 2, no. 9, p. e0000324, 2023.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 77-85. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2742

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

AN ENHANCED MULTIMODAL APPROACH FOR SENTIMENT ANALYSIS USING DEEP LEARNING TECHNIQUES

VinithaV[1]* and S.K. ManjuBargavi[2]

[1]Research scholar, DepartmentofComputerScienceandIT,Jain(Deemed-to-be-University),Bengaluru,India
[2]Professor, Department of Computer Science and IT, Jain(Deemed-to-be-University),Bengaluru,India

[1]https://orcid.org/0009-0006-4586-5923, [2]https://orcid.org/0000-0001-7112-5810

Email: * vinithafive@gmail.com, b.manju@jainuniversity.ac.in

ARTICLE INFO

Article History
Received: September 25, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Multimodal,
Hybrid remoraoptimization,
Depression detection,
Sentiment analysis,
Deep learning.

ABSTRACT

Social media has become a prominent medium through which individuals express their perspectives, sentiments, and experiences, frequently exhibiting apparent indications of mental health conditions such as depression. The identification of depression through nonverbal behaviour has garnered considerable interest. Implementing previous research concerning the detection of depression within real-world scenarios presents challenges, primarily because the research predominantly concentrated on identifying depressive individuals within controlled laboratory settings. Depression affects millions worldwide therefore, early and accurate detection is critical for prompt intervention. Conventional diagnostic methodologies are often subjective in nature and require considerable time to implement. This study introduces a hybrid remora-optimized multimodal deep learning (HRO-MDL) approach to incorporate multi-modal data such as audio, video, and text for enhanced performance in the assessment of depressive emotional states and sentiments. The proposed framework accurately predicts sentiment across modalities by integrating feature extraction from various modalities using a deep convolutional neural network optimized with hybrid optimization. The model's performance was assessed using the D-Vlog and Depression Cleaned Reddit datasets, demonstrating notable accuracy levels 94.00% for text data, 92.00% for audio data, and 93.00% for video data. In contrast to traditional methodologies, the presented work exhibits enhanced efficiency and underscores its suitability for diverse multimodal sentiment analysis applications.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The significant influence of mental health on both individual and societal well-being has resulted in increased focus on this issue among diverse populations. Mental health disorders, such as depression, are commonly occurring conditions that drastically impact millions of people worldwide. The disease results with constant pain, deep despair and considerably impaired functional capacity that leads to serious limitation in daily life [1]. Current method for diagnosing depression is primarily self-reported questionnaires, interviews, and clinical measurements. However, these approaches can be costly, time-consuming and subject to human perception bias [2]. To assess people's emotional states, clinicians are using modality data including text, audio and visual [3]. An accurate and real-time detection can be realized if the analysis is considering multiple depression data sources including text, voice, and visual to learn depression more explicitly [4]. The combination of multiple modalities in such systems is helpful in improving result of confidence, as it uses the strengths of different signals that reduces the chances of signal failure and misinterpretation. The approach maintains the continuous monitoring with increased precision with effect of early detection [5] and even in the analysis of large data such as text, audio and video data, there are various more advanced models (such as CNN's, RNN's and transformers) showing increased effectiveness. Medical professionals can leverage these models to identify signs of depression, which may present as subtle behaviours that includes inclination toward the condition. The methods for detecting depression support the advancement of deep learning technologies, enabling precise and

early identification of depressive symptoms [6]. Current state-of-the-art approaches, including CNNs, RNNs, and transformers, demonstrate robust performance in analysing multimodal data comprising text, audio, and video that involves substantial textual content. The model assists healthcare providers in identifying the symptoms of depression and detects subtle behavioural changes indicative of a tendency toward depressive patterns [7]. The primary challenge is to improve model efficiency while maintaining generalizability across diverse populations with varying linguistic and socioeconomic attributes [8].

This method offers a significant degree of accuracy independently however, the necessity for these models to be resilient and effective leads to overfitting due to insufficient real-world applicability [9]. This study explores a multi-modal deep learning approach incorporating hybrid remora- Integrated Grey Wolf Optimizationtechniques. This method is structured to preprocess various types of input data (text, audio, video) through multiple stages, including data preprocessing and subsequently proceeding to multimodal feature extraction. DenseNet-201 classification is utilized to predict emotional states of depression and sentiments, as it is well-suited for diverse input modalities. Additionally, this optimization technique is employed to refine the models and improve system performance. The hybrid optimization enhances system accuracy, rendering it robust for practical applications. Below is the major contribution of the proposed approach:

- To develop a multi-modal deep learning approach with hybrid remora optimization and to predict the emotional state of depression and sentiments across various modalities.
- The proposed model processes the sentiment of each modality (text, acoustic, visual) by modelling the verbal and non-verbal behaviour over the modalities.
- The Hybrid Remora-Integrated Grey Wolf Optimizationwith the DenseNet-201 approach is utilized to compare with current baseline models and achieve the best feature set with the highest accuracy.

II. LITERATURE REVIEW

Depression is one of the most prevalent and severe mental health disorders identified on social media. The development of an attention-based approach for detecting depression. This model emphasized the most important features of the input data and increased its interpretability. The observed overfitting was due to the model's increasing complexity [10]. Spatial attention networks are created to detect depression in facial images. The approach demonstrated significant efficacy and improved the detection parameters. The dataset for this model was significantly smaller and more complex [11]. Four hybrid deep learning models were created specifically to identify symptoms of melancholy. This method may function as an efficient instrument for detecting symptoms of depression on social media platforms, owing to its improved accuracy and relevance to real-world textual data. This method largely analysed textual data, overlooking potentially beneficial multimedia data such as images or video snippets, that could offer a more comprehensive understanding of depression [12], [13]. They devised an attention technique that amalgamated voice and text analysis to detect depression. This method derives contextual dependencies from both prior and later segments of the sequence by analysing data in both forward and reverse directions.

The bidirectional data processing incurred computational costs and led to an overfitting problem [14]. A fusion fuzzy model was used to identify depression from facial expressions. For visual data, this approach demonstrated strong feature extraction capabilities, but it becomes challenging when working with large datasets [15]. The deep learning model performed better at detecting depression in terms of accuracy rate and using data from Twitter. The model is susceptible to overfitting to the specific dataset, which could restrict its effectiveness when applied to newly obtained, unseen data [16]. A method for identifying depression and capturing local dependencies as well as sequential relationships within the text was introduced using deep learning. By capturing the semantic substance of sentences not explicitly present in the lexicon, this approach enhanced the model's ability to interpret subtle verbal indicators of depression. It required a substantial amount of computational power, especially when handling big datasets [17].

Experimental outcomes emphasize the advantages of integrating deep learning with multi-aspect data, indicating that it exceeds several established and robust baseline methods [18]. The Particle Swarm Optimization algorithm was developed for the diagnosis of depression. This approach enhances both precision and the rate of convergence. The accuracy of the model is significantly influenced by the quality and diversity of the dataset. Imbalanced datasets can negatively impact the model's performance, resulting in inaccurate melancholy detection, despite the attention given to recognizing melancholy in social media [19-21]. Their research has enhanced the model's interpretability and demonstrated the significance of features within the input data. Due to the increased complexity of the models, overfitting occurred. For the purpose of detecting depression in facial images, this study employed spatial attention networks [22], [23]. The primary benefits of this methodology were its enhanced performance characteristics and the improved pertinence of the generated features for detection purposes. The resultant model was both decreased in quantity and increased in complexity [24].

III. PROPOSED METHODOLOGY

This study focuses on the Hybrid RemoraOptimizedMultimodal Deep Learning (HRO-MDL) approach which is proposed to predict the emotional state of depression and sentiment detection. Depression detection through the integration of multi-modal data (video, audio, and text) utilizing different methods. It encompasses a variety of phases, including data preprocessing, parameter tuning, classification, and optimization, as illustrated in Figure 1.

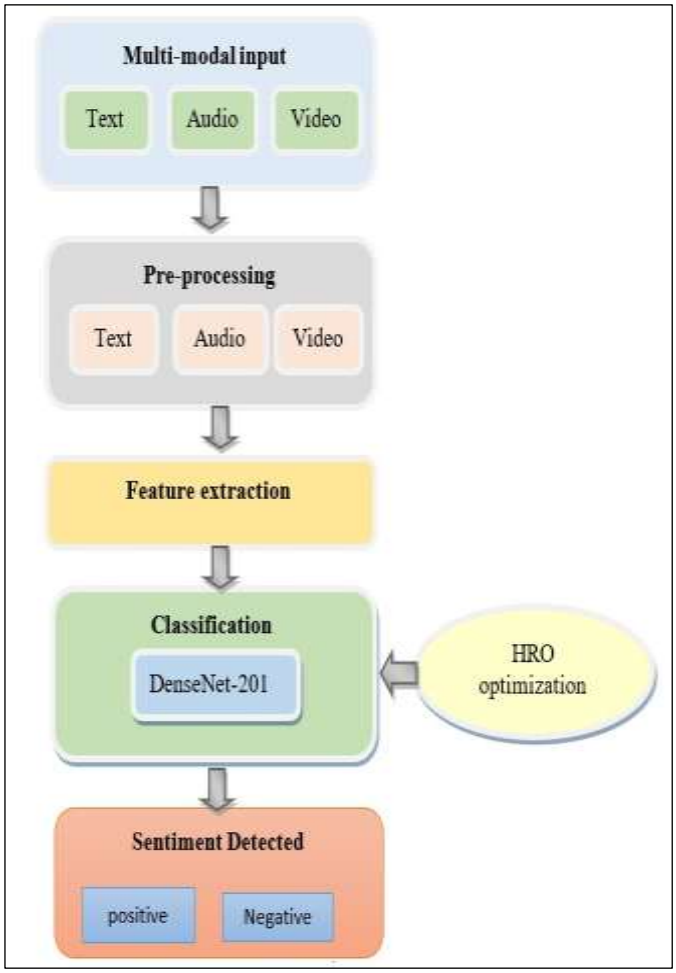


Figure 1: Proposed Multi-Modal Framework.
Source: Authors, (2026).

III.1 DATA COLLECTION

The input source audio and video data are collected from the D-Vlog dataset. The textual dataset is taken from Cleaned Reddit dataset. The audio features that supply important cues to diagnose depression, such as facial expressions, speech patterns, and emotional response signals, are also taken from the D-Vlog dataset, which also includes video logs of human life experiences. It consists of depression-related linguistic patterns as well as sentiment indicators and psychological markers for depression assessment. The mathematical representation of multi-modal input is represented as follows.

The D-Vlog dataset is the source of the input audio and video data. Cleaned Reddit depression dataset used for textual data. Also extracted D-Vlog datasetincludes video records of human life experiences, audio elements that provide crucial clues to diagnose depression, such as speech patterns, facial expressions, and emotional response signals. It includes sentiment indicators, psychological markers for depression assessment, and linguistic patterns associated with depression. Multi-modal input can be represented mathematically as follows:

$$Y = \{Y_T, Y_A, Y_V\} \tag{1}$$

Whereas, audio and video data from the D-Vlog dataset is represented byY_A,Y_V, and text data denoted as Y_T from the Reddit Cleaned depression dataset.

III.2 PRE PROCESSING

Preprocessing of audio, video, and textual data is utilized to extract multimodal features for sentiment prediction and depression detection. Speech is pre emphasized, framed, and converted to the Mel Frequency Cepstral Coefficients (MFCC), which are used to measure speechand ambient clues. Videos are converted into frames, resized, normalized, and spatially featured positions captured by the CNNs and temporally motion patterns learned by optical flow. Text data preprocessing, involving tokenization, lemmatization, stop word removal, and embedding generation, is conducted to preserve the semantics and context of the text data for analysis. Let the text input be expressed in Equation 2.

$$Y_T = \{m_1, m_2, \ldots . m_n\} \tag{2}$$

Where, mi represents the individual sentence

III.3 FEATURE EXTRACTION

The Audio, visual, and text are all used to extract multimodal features to model sentiment detection using the Dvlog dataset and Reddit dataset. Then texts are pre-processed through tokenization and cleaning and are transformed into contextual embeddings based on XLNet, which can obtain deep semantics and syntax relations. Also, the audio streams are pre emphasized, framed, and converted to Mel Frequency Cepstral Coefficients (MFCC) suitable for time sequence modelling with XLNet on temporal acoustic patterns. Videos are split into frames and resized, and then normalize, spatial features are extracted with CNNs (e.g., DenseNet), and temporal dependencies models such as XLNet transformers with optical flow for motion dynamics. Given a sequence of input tokens $\boldsymbol{y} = [\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots \boldsymbol{y_T}]$, XLNet models the joint probability of a given input sequence of tokens as per Equation 3.

$$P(y) = \prod_{t=1}^{T} P(y_t | y_{<t}) \tag{3}$$

The set of tokens that have been proceeding in the permutation is depicted. XLNet's ability to analyse all sequence permutations, which makes each token accessible to all other tokens, can be interpreted to serve as an alternative to the traditional prediction approach systems, where tokens are only conditioned on a fixed left-to-right context through an autoregressive model. As a result, it achieves bidirectional dependency understanding. The anticipated log likelihood is optimized by the model for each possible sequence of manufacturing order permutations as per Equation 4

$$\hat{\theta} = arg\, \max_{\theta} E_{Z\sim\Im}\left[\sum_{t=1}^{T} log\, P\left(y_{Z(t)} | y_{z(<t)}\right)\right] \tag{4}$$

III.4. CLASSIFICATION

Once that features are gathered, they are used as input to the DenseNet-201 deep learning model for classification, which includes efficient dense connections elements that includes feature reuse but also improves gradient flow. The applicability of the model parameters is enhanced through Hybrid Remora-Integrated Grey Wolf Optimization with the faster converging speed, and the highest classification accuracy. The collaboration of extraction methods and optimized classification techniques leads to solid outcomes in both depression evaluations and sentiment analysis processes. Each layer of Dense Net- 201 is based on the dense block concept, Ds receive the output of all preceding layers 〖D〗_1,D_2,....D_(s-1). The output of the s-th layer in a dense block can be represented as follows for a particular input feature map as per Equation 5.

$$Y_s = H_s([Y_0, Y_1, \ldots Y_{s-1}]) \tag{5}$$

Where, the output of the s-th layer is indicated by Y_s, while the function is shown by $H_s$applied at layer s,$[Y_0, Y_1, \ldots Y_{s-1}]$ representing the concatenation of all previous layer output. DenseNet-201 runs several dense blocks, transition layers, and a fully linked layer for classification to provide its final output. The SoftMax function is used to generate the output probability vector at the end of the classification process represented in Equation 6.

$$P(x|Y) = soft\, max(WY_T + b) \tag{6}$$

Where Y_T denotes the final feature map after being passed through all layers, W and b are weights and bias of the fully connected layer,$P(x|Y)$ denotes the probability distribution over classes after the softmax activation. The usage of a SoftMax function for classification and DenseNet-201's dense connection are reflected in the above formulation.

III.5 HYBRID REMORA OPTIMIZATION ALGORITHM

It is a kind of population based metaheuristic optimization algorithm that imitates hierarchical hunting strategy of Grey Wolf Optimization (GWO) for group hunting and embodies the symbiotic follow-and-adapt behaviour of remora fish in oceanic ecosystems. Within this hybrid framework, the leadership hierarchy of Grey Wolf serves as the primary search approach, steering global exploration of solution space. The follow-and-adapt mechanism as a biologically inspired approach underlies remora integration strategyand serves as an effective mechanism for both exploration and exploitation balance and thus enhancing convergence speed and solution quality. The positions of remora agents will be updated in each iteration according to the associated position update formulation defined in Equation (7) during the remora adaptation phase.

$$Y_i(t+1) = Y_{best} + \beta(Y_{best} - Y_i(t)) \tag{7}$$

The new position of remora i is represented as Yi(t+1), the optimal solution is indicated as Y_best , and the adaptive parameter governing fine-tuning is denoted as β . The Eat Thoughtfully Phase enables remoras to leverage possible solutions, resulting in accelerated convergence and enhanced accuracy, particularly in complex optimization challenges such as hyperparameter tuning in deep learning. However, it achieves better exploration and exploitation balance through an integration with RSO. The purpose behind this modification serves to stop early convergence while simultaneously enabling better global exploration and effective optimization of deep learning models. This proposed method enables superior performance with optimal accuracy and robustness.

IV. RESULTS AND DISCUSSION

This section focuses on assessing the performance of the proposed system and comparing it against baseline approaches using standard benchmark datasets. To make the model more robust, 2 standard datasets (DVLOG, CLEANED REDDIT), are used to integrate to form a hybrid model.

IV.1. DATASET DESCRIPTION

- D-Vlog Dataset [19]: It consists of 961 YouTube vlogs (160 hours) used for depression detection using non-verbal behaviour. The dataset contains 816 subjects (406 controls, 555 sufferers of depression) with audio and visual features saved as NumPy files. This supports the automated depression detection in research with AI-based analysis of facial expressions, gestures, and speech patterns as they naturally occur in everyday life.
- DepressionReddit Dataset: This dataset contains 7,000 labelled text data that are collected from depression related subreddits. Text data reflects real emotional and mental health conditions through self-expressed content. It is valuable for sentiment and depression detection, offering linguistic markers that help identify psychological distress using natural language processing techniques.

IV.2. EXPERIMENTAL SETUP

In the context of this study, where training, testing is done, after was 70% data trained and other thirty percent is imposed to test. The study done on a 2.4 GHz Intel(R) processor with 16 gigabytes of random-access memory (RAM).

IV.3. EVALUATION METRICS

Several evaluation metrics were used to assess the global classifying efficiency. Table 1 was used to evaluate the proposed approach for different evaluation criteria to analyse the sentiments with various performance metrics as shown in Table 1 and Fig.4.

IV.4. CONFUSION MATRIX HEATMAP

It denotes the evaluation of a classification model. It shows real classifications versus predictions in a matrix arrangement across modalities shown in below Figure 2, the confusion matrix shows how well system predict accurate and inaccurate predictions across depression vs normal classes over different attributes. Darker cells correspond to high-accuracy predictions, and lighter cells are misclassifications. In Figure 3. Shows Training and validation accuracy and loss curve for various modalities, Figure 4 shows ROC Curve across thresholds on classifier performance, AUC of 0.95–0.98 indicates that our sentiment classification is highly sensitive and specific.

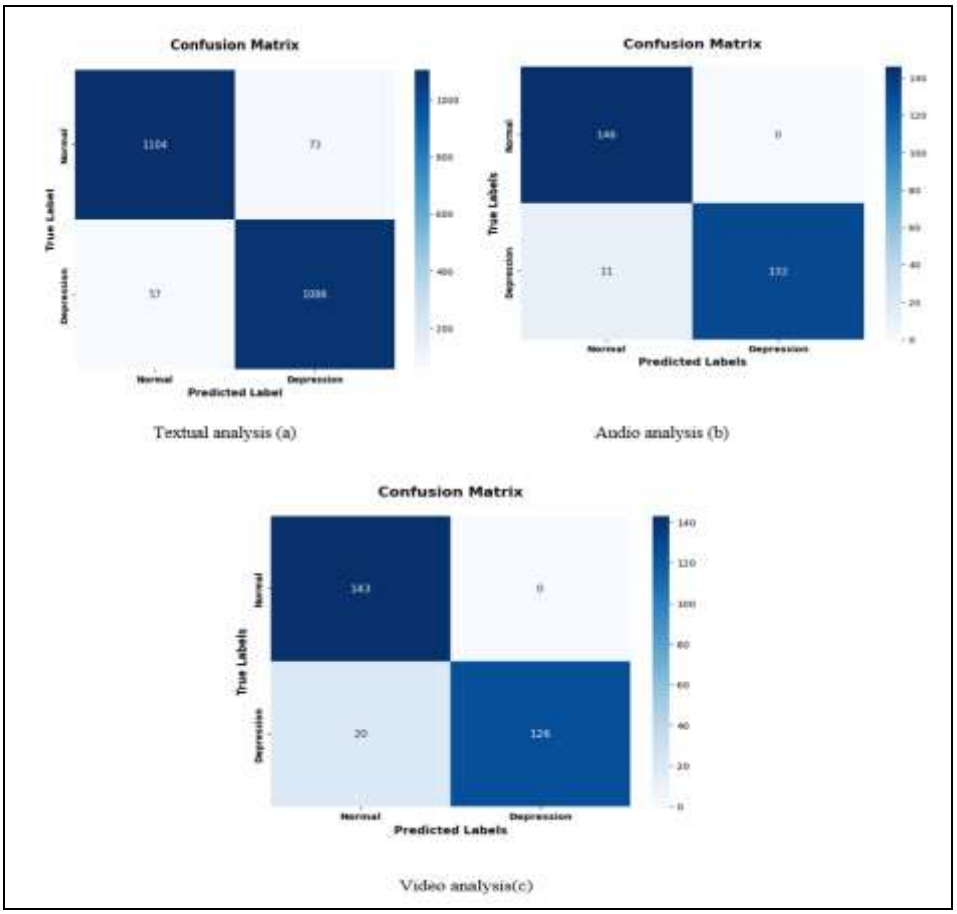


Figure 2: Confusion matrix heat map for text, audio, video modalities.
Source: Authors, (2026).

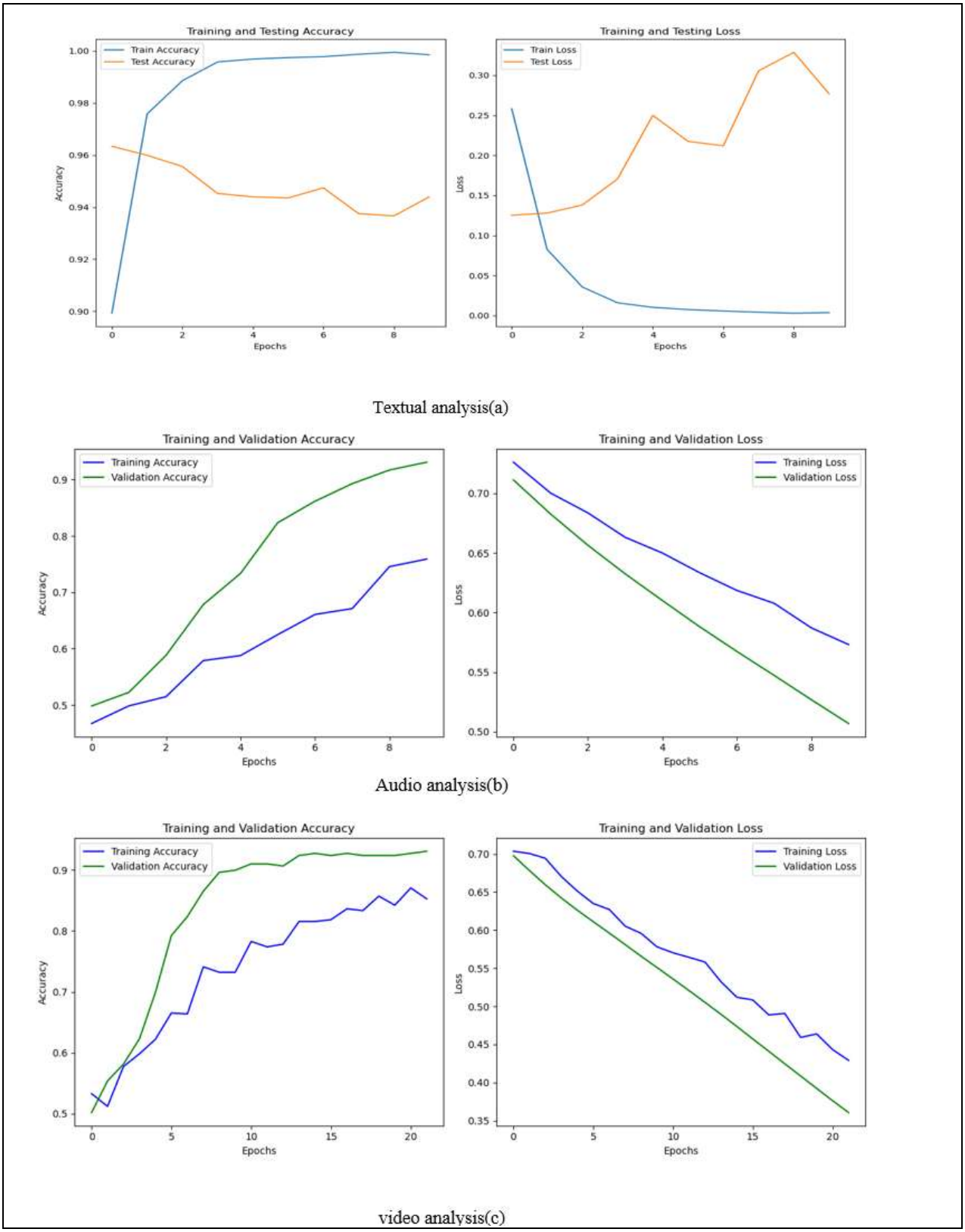


Figure 3: Training and Validation Accuracy and loss acrossmodalities.
Source: Authors, (2026).

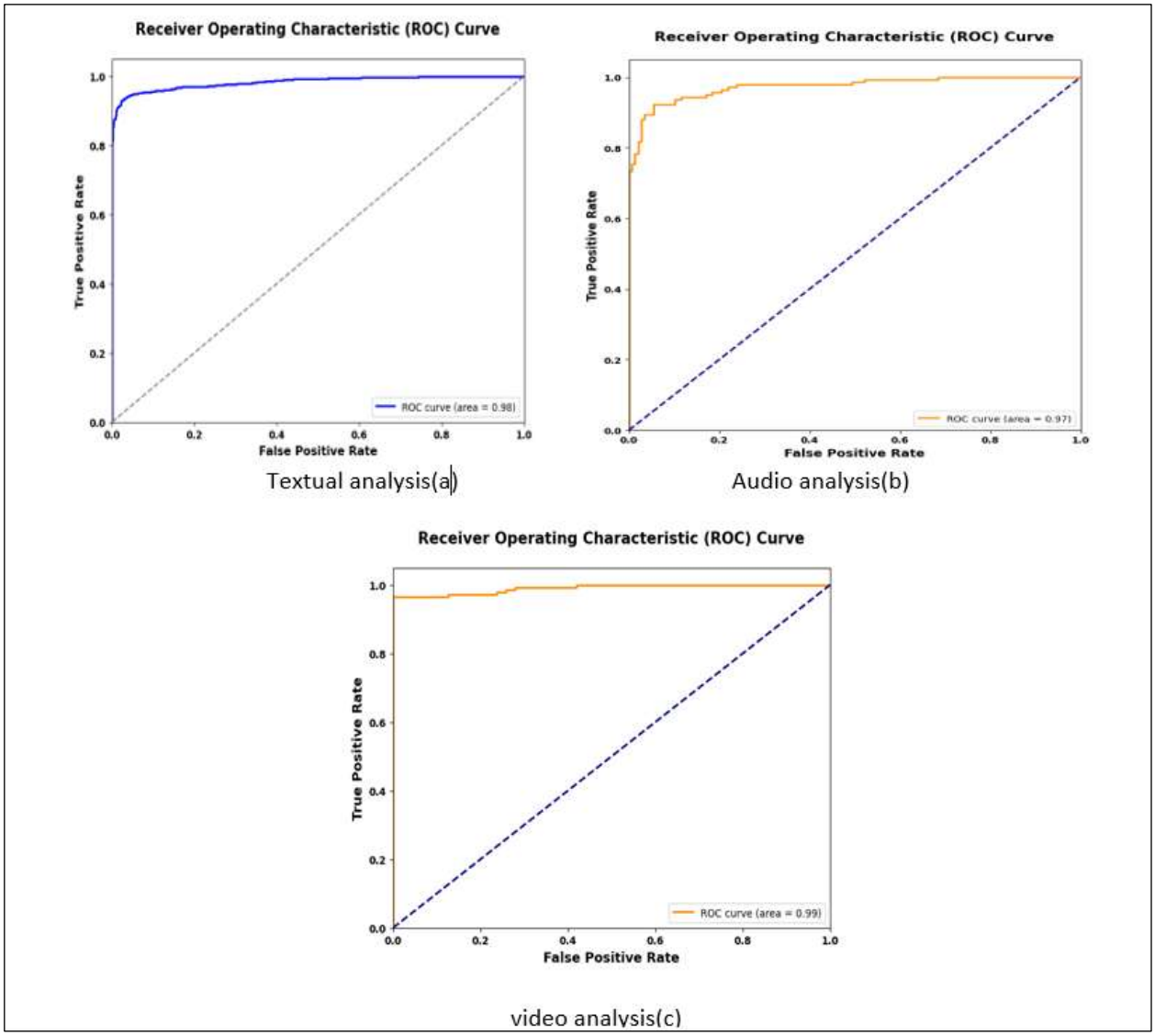


Figure 4: ROC for various modalities.
Source: Authors, (2026).

IV.5. COMPARATIVE STUDY WITH VARIOUS BASELINE MODELS

The proposed depression detection system is a webbasedapplication that contains multiple interactive modulestoexecute regular mental health assessments. The Base line models such as Tensor fusion Network (TFN), Depression Detector, Time-Aware Attention Multimodal Fusion Network (TAMFN), STST Spatiotemporal Semantic Transformer (STST) were tested over problems in depression detection like inability of handling sequential data, high computation complexity and slower execution on very large datasets shown by Table 1 and Figure 5. However, these methods consistently underperform due to issues related to hyperparameters, overfitting, and convergence. Issues can be addressed through integrationof Hybrid Remora-Integrated Grey Wolf Optimization, usingDenseNet-201 for classification and Adam optimizer was used for fine tuning to improve the performance. It also provides a real time facial landmark detection, contributes to depression assessment and there other several use cases. The different baseline models are discussed below.

1. **TFN (Tensor Fusion Network):** This is a multimodal sentiment baseline that models unimodal, bimodal and trimodal interactions explicitly through tensor outer-products, it's strong but it has many parameters, it works poorly when the data is very sparse.
2. **Depression Detector:** Audio + Text bi-modal depression detector with GRU/BiLSTMmodel for temporal modeling, which is strong base-midline for clinical/interview-style data.
3. **TAMFN:** Time-Aware Attention Multimodal Fusion Network is used for learning the temporal salience and cross-modal fusion for depression detection which deals with the temporal handling issue better than TFN.
4. **STST:** Spatiotemporal Semantic Transformer uses attention over visual/text/audio streams with semantic alignment, improved long-range temporal modeling and fusion.
5. **TVLT**:(textless pretraining) token-based vision-language transformer, re-purposed for cross-modal understanding and strong generic prior, moderate domain gap without task-specific pretraining.

Table 1: Comparison across various baseline models.

Model Type	Precision	Recall	F1-Score	Acc
TFN [18]	61.39%	62.26%	61.00%	61.55%
Depression Detector [19]	65.40%	65.57%	63.50%	64.82%
TAMFN [20]	66.02%	66.50%	65.82%	66.11%
STST [21]	72.50%	77.67%	75.00%	75.06%
TVLT [22]	67.30%	68.03%	67.80%	67.80%
Proposed Model				
Text Analysis	94.00%	95.00%	94.00%	94.00%
Audio Analysis	93.00%	91.00%	92.00%	92.00%
Video Analysis	94.00%	92.00%	93.00%	93.00%

Source: Authors, (2026).

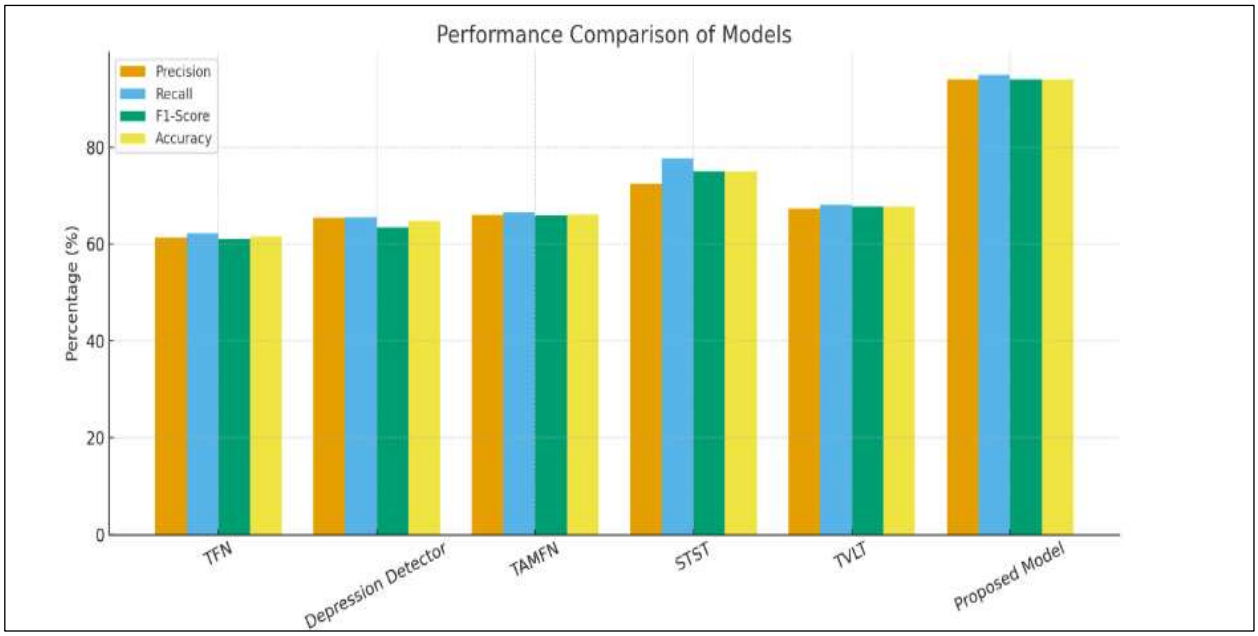


Figure 5: Comparison graph with various Baseline Models.
Source: Authors, (2026).

V. CONCLUSION

Despite the field of deep learning is constantly expanding, there is still a lot of work accessible for depression identification. This study proposes to use multimodal data (text, video, and audio) that is effectively incorporated to detect emotional states of depression and classify sentiment. High-level features between modalities can then be extracted using feature extraction techniques. The Dense Net-201 model is then used to achieve an accurate classification, and a Hybrid Remora-Integrated Grey Wolf OptimizationAlgorithm has been introduced to further improve the classification accuracy. Initially, it preprocesses the various input data so that it can support text, audio, and video data separately with the available data. Thus, it has been demonstrated that this suggested method outperforms current methods in the comparative analysis, with greater accuracy rates for text data with 94%, audio data is 92%, and video data is 93%. The integration of real-time sensor physiology will be the focus of future research in order to create a more interpretable visualization of the information that the mental health professionals are actually acquiring from the decisions being performed.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Vinitha V, S.K. Manju bargavi.
Methodology: Vinitha V, S.K. Manju bargavi.
Investigation: Vinitha V, S.K. Manju bargavi.
Discussion of results: Vinitha V, S.K. Manju bargavi.
Writing –Original Draft: Vinitha V.
Writing –Review and Editing: Vinitha V, S.K. Manju bargavi.
Resources: Vinitha V, S.K. Manju bargavi.
Supervision: Vinitha V, S.K. Manju bargavi.
Approval of the final text: Vinitha V, S.K. Manju bargavi.

VII. ACKNOWLEDGMENTS

The authors would like to extend their heartfelt thanks to the Department of Computer science and IT, Jain University, Bengaluru, for their support of this work.

VIII. REFERENCES

[1] S. Khan and S. Alqahtani, “Hybrid machine learning models to detect signs of depression,” Multimedia Tools and Applications, vol. 83, no. 13, pp. 38819–38837, 2024.

[2] K. Jawad et al., “Novel cuckoo search-based metaheuristic approach for deep learning prediction of depression,” Applied Sciences, vol. 13, no. 9, p. 5322, 2023.

[3] N. Marriwala and D. Chaudhary, “A hybrid model for depression detection using deep learning,” Measurement: Sensors, vol. 25, p. 100587, 2023.

[4] A. S. Rajawat et al., “Fusion fuzzy logic and deep learning for depression detection using facial expressions,” Procedia Computer Science, vol. 218, pp. 2795–2805, 2023.

[5] H. A. Sanghvi et al., “A deep learning approach for classification of COVID and pneumonia using DenseNet-201,” International Journal of Imaging Systems and Technology, vol. 33, no. 1, pp. 18–38, 2023.

[6] Y. Pan et al., “Spatial–temporal attention network for depression recognition from facial videos,” Expert Systems with Applications, vol. 237, p. 121410, 2024.

[7] W. Zhang, J. Xie, Z. Zhang, and X. Liu, “Depression detection using digital traces on social media: A knowledge-aware deep learning approach,” Journal of Management Information Systems, vol. 41, no. 2, pp. 546–580, 2024.

[8] D. K. Saha et al., “Ensemble of hybrid model based technique for early detecting of depression based on SVM and neural networks,” Scientific Reports, vol. 14, p. 25470, 2024.

[9] N. K. Iyortsuun et al., “Additive cross-modal attention network (ACMA) for depression detection based on audio and textual features,” IEEE Access, vol. 12, pp. 20479–20489, 2024.

[10] A. S. Rajawat et al., “Fusion fuzzy logic and deep learning for depression detection using facial expressions,” Procedia Computer Science, vol. 218, pp. 2795–2805, 2023.

[11] D. S. Khafaga et al., “Deep learning for depression detection using Twitter data,” Intelligent Automation & Soft Computing, vol. 36, no. 2, pp. 1301–1313, 2023.

[12] V. Tejaswini, K. S. Babu, and B. Sahoo, “Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model,” ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 23, no. 1, pp. 1–20, 2024.

[13] A. F. Dharma, “A deep learning using DenseNet201 to detect masked or non-masked face,” 2021.

[14] H. Zogan et al., “Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media,” World Wide Web, vol. 25, no. 1, pp. 281–304, 2022.

[15] J. P. Thekkekara, S. Yongchareon, and V. Liesaputra, “An attention-based CNN-BiLSTM model for depression detection on social media text,” Expert Systems with Applications, vol. 249, p. 123834, 2024.

[16] Y. Pan et al., “Spatial–temporal attention network for depression recognition from facial videos,” Expert Systems with Applications, vol. 237, p. 121410, 2024.

[17] M. Kächele, M. Schels, C. Thiel, and F. Schwenker, “Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression,” in Proc. 6th Int. Workshop on Audio/Visual Emotion Challenge, pp. 11–18, 2016.
[18] J. Yoon, C. Kang, S. Kim, and J. Han, “D-vlog: Multimodal vlog dataset for depression detection,” in Proc. AAAI Conf. Artif. Intell., 2022.

[19] L. Zhou et al., “TAMFN: Time-aware attention multimodal fusion network for depression detection,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 669–679, 2022.

[20] Y. Tao et al., “Depressive semantic awareness from vlog facial and vocal streams via spatio-temporal transformer,” Digital Communications and Networks, vol. 10, no. 3, pp. 577–585, 2024.

[21] P. Moon and P. Bhattacharyya, “We care: Multimodal depression detection and knowledge infused mental health therapeutic response generation,” arXiv preprint arXiv:2406.10561, 2024.

[22] K. S. Seby, M. Elamparithi, and V. Anuratha, “A hybrid attention-based deep learning system for suicidal ideation detection in social media,” 2024.

[23] E. Yeskuatov, S. L. Chua, and L. K. Foo, “Detecting suicidal ideations in online forums with textual and psycholinguistic features,” Applied Sciences, vol. 14, no. 21, p. 9911, 2024.

[24] S. L. Mirtaheri, S. Greco, and R. Shahbazian, “A self-attention TCN-based model for suicidal ideation detection from social media posts,” Expert Systems with Applications, vol. 255, p. 124855, 2024.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 86-103. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2744

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

OPTIMIZING OVERLAP IN TREE-BASED INDEXING STRUCTURES FOR ENHANCED K-NN SEARCH EFFICIENCY

Ala-Eddine Benrazek*[1], Ibtissem Kemouguette[2], Zineddine Kouahla[3], Brahim Farou[4] and Hamid Seridi[5]

[1] Department of Computer Science, CSAAIL Laboratory, Ziane Achour University, P.O BOX 17000, Djelfa, Algeria.
[1, 2, 3,4,5] Department of Computer Science, LabSTIC Laboratory, 8 mai 1945 University, P.O BOX 24000, Guelma, Algeria.

[1] http://orcid.org/0000-0003-2182-6664, [2]http://orcid.org/0009-0006-6994-0158, [3]http://orcid.org/0000-0001-8105-9810, [4]http://orcid.org/0000-0002-1609-6006, [5]http://orcid.org/0000-0002-0236-8541

Email: * alaeddine.benrazek@univ-djelfa.dz, kemouguetteibtissem@gmail.com, kouahla.zineddine@univ-guelma.dz, farou.brahim@univ-guelma.dz, seridi.hamid@univ-guelma.dz

ARTICLE INFO

Article History
Received: September 26, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Internet of Things (IoT),
Data Indexing,
k-NN Search,
Overlap Optimization,
Tree-Based Indexing.

ABSTRACT

The proliferation of interconnected devices in the Internet of Things (IoT) has led to an exponential increase in data, commonly known as Big IoT Data. Efficient retrieval of this heterogeneous data demands a robust indexing mechanism for effective organization. However, a significant challenge remains the overlap in data space partitions during index construction. This overlap increases node access during search and retrieval, resulting in higher resource consumption and performance bottlenecks and impedes system scalability. To address this issue, we propose three innovative heuristics designed to quantify and strategically reduce data space partition overlap. The volume-based method (VBM) offers a detailed assessment by calculating the intersection volume between partitions, providing deeper insights into spatial relationships. The distance-based method (DBM) enhances efficiency by using the distance between partition centers and radii to evaluate overlap, offering a streamlined yet accurate approach. Finally, the object-based method (OBM) provides a practical solution by counting objects across multiple partitions, delivering an intuitive understanding of data space dynamics. Experimental results demonstrate the effectiveness of these methods in reducing search time, underscoring their potential to improve data space partitioning and enhance overall system performance.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The Internet of Things (IoT) has rapidly emerged as one of the most transformative technologies of the 21st century. This paradigm is characterized by intelligent, self-configuring devices capable of communicating with one another through a global network infrastructure [1]. Since Kevin Ashton introduced the concept of IoT in 1999 [2], the number of connected devices has skyrocketed from 8 billion in 2018 to an estimated 75 billion by 2025 [3]. Alongside this growth, the volume of generated data has increased from 33 ZB in 2018 to a projected 175 ZB by 2025 [4], marking the dawn of the big IoT data era [5](see Figure 1).

This exponential growth of data in IoT-based applications offers significant opportunities for value creation but also highlights the urgent for efficient data storage and retrieval mechanisms [6]. Similarity search-based data retrieval is widely adopted for its ability to retrieve complex data based on intrinsic characteristics [7]. However, its effectiveness is often limited by high computational costs, especially in big data settings. To effectively leverage this technique and avoid the brute force sequential scan, robust indexing methods are essential for organizing large datasets in a manner that facilitates the efficient retrieval of similar data [8], [9]. In this context, the concept of metric space plays a crucial role, allowing objects to be mapped into a domain where they can be compared using a distance function [10].

Typically, indexing techniques within this space divide data into hierarchical regions centered around selected representative objects, known as pivots, each with a defined covering radius. While each region is associated with a pivot, every object within the region must be explicitly assigned.

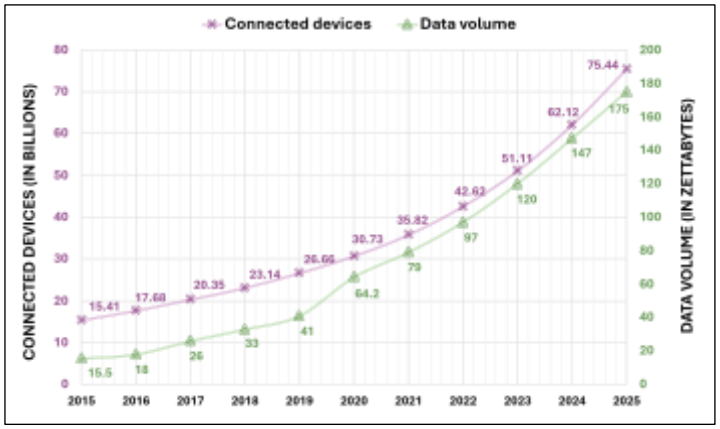


Figure 1: Growth of Connected Devices and Data Volume (2015-2025).
Source: Authors, (2026).

Several metric index methods, such as the M-tree [11], Slim-Tree [12], and more recently, the BCCF-tree [6] and its extension, the B3CF-tree [13], have been developed to enhance similarity searches in metric spaces. Despite these advancements, the issue of space overlap in index structures persists, leading to increased node accesses, longer search times, and higher computational costs—sometimes even exceeding those of sequential scans. Rather than resorting to solutions that eliminate all elements within the overlap area or minimize partitioning radii—both of which can lead to information loss—there is a critical need for a new mechanism to more effectively quantify, measure, and address the overlap problem. Such a mechanism should incorporate advanced metrics for accurately detecting overlaps and sophisticated methods for evaluating them during the partitioning process. To bridge this significant research gap, we propose three innovative heuristics designed to quantify and strategically manage data space partition overlap:

- Volume-Based Method (VBM): This heuristic provides a detailed measure of overlap by calculating the intersection volume between partitions. It offers a nuanced understanding of spatial relationships, enabling reliable quantification of overlap areas and optimizing partition boundaries.
- Distance-Based Method (DBM): This approach evaluates overlap by analyzing the distance between partition centers and their radii. By focusing on the geometric relationships between partitions, this method ensures a computationally efficient process without sacrificing accuracy.
- Object-Based Method (OBM): Notable for its practicality, this heuristic uses the count of objects shared between partitions as a straightforward yet insightful indicator of overlap. It facilitates a more intuitive interpretation of data space dynamics by directly linking overlap to data distribution.

This paper is organized as follows: **Section 2** offers an overview of the essential background information. **Section 3** reviews relevant existing works. **Section 4** presents the proposed heuristics in detail. **Section 5** analyses the comparative experimental results. Finally, **Section 6** concludes the study and outlines future research directions.

II. THEORETICAL BACKGROUND

The metric space is a mathematical structure, introduced as a universal abstraction for data [14], excels by relying solely on a distance function that satisfies the triangle inequality, thereby eliminating the need for specific content or intrinsic structure [15], [16]. This versatility allows it to effectively index a wide variety of data types seamlessly.

Definition 1 (Metric Space). *A metric space $\mathcal{M}$ is defined as a pair $(\mathcal{S}, d)$, where $\mathcal{S}$ denotes a domain of objects, and $d: \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}^+$ represents a distance function to quantify the "similarity" between objects within $\mathcal{S}$. This distance function typically satisfies four distinct properties:*

- $\forall (x, y) \in \mathcal{S}, d(x, y) \geq 0$ $\quad (p1: non-negativity)$
- $\forall (x, y) \in \mathcal{S}, d(x, y) \geq d(y, x)$ $\quad (p2: symmetry)$
- $\forall (x, y) \in \mathcal{S}, d(x, y) = 0 \iff x = y$ $\quad (p3: identity)$
- $\forall (x, y, z) \in \mathcal{S}, d(x, y) + d(y, z) \geq d(x, z)$ $\quad (p4: triangle\ inequality)$

According to [17], the problem of indexing metric spaces can be formally stated as follows:

Definition 2 (Problem of Data Indexing). *Let $\mathcal{M} = (\mathcal{S}, d)$ be a metric space. Consider a subset $\mathcal{X} \subseteq \mathcal{S}$, representing a set of objects, commonly referred to as the dataset, within which we perform searches. The problem of indexing metric spaces involves determining how to preprocess or structure the objects in $\mathcal{X}$ so that similarity queries can be answered efficiently.*

As mentioned in Section 1, indexing techniques in this context typically involve partitioning a set $\mathcal{X} \subseteq \mathcal{S}$ of objects in the metric space $\mathcal{M} = (\mathcal{S}, d)$ into hierarchical regions centered around pivots using various partitioning techniques. One widely used method is Generalized Hyperplane (GH) partitioning, which is commonly employed in these indexes [8], [16]. In GH partitioning, two specific objects $(p_1, p_2) \in \mathcal{X}$ are chosen arbitrarily or specifically. The dataset $\mathcal{X}$ is then divided into subsets $\mathcal{X}_1$ and $\mathcal{X}_2$ based on Definition (3) and as illustrated in Figure 2.

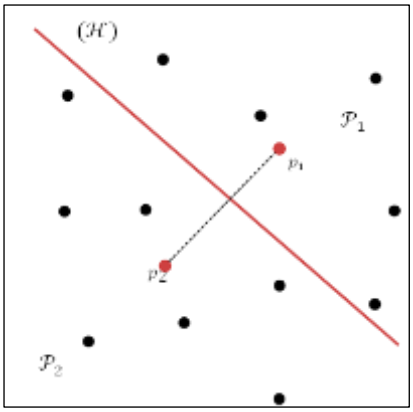


Figure 2: GH partitioning $\mathcal{H}(\mathcal{S}, d, p_1, p_2)$.
Source: Authors, (2026).

Definition 3 (Generalized Hyperplane Partitioning). *Let $\mathcal{M} = (\mathcal{S}, d)$ be a metric space, where $\mathcal{X} \subseteq \mathcal{S}$ represents the set of objects to be indexed. Assume $(p_1, p_2) \in \mathcal{X}$ are two pivot objects such that $d(p_1, p_2) > \mathcal{X}$. The generalized hyperplane $\mathcal{H}(\mathcal{S}, d, p_1, p_2) = \{o \in \mathcal{X} : d(p_1, o) = d(p_2, o)\}$ partitions the space into two regions as follows:*

$$\begin{cases} \mathcal{P}_1 = \{o \in \mathcal{X} : d(p_1, o) \leq d(p_2, o)\} \\ \mathcal{P}_2 = \{o \in \mathcal{X} : d(p_1, o) \geq d(p_2, o)\} \end{cases}$$

In metric space theory, a similarity query is defined by a query object q and a specified proximity criterion, typically expressed as a distance. The objective of such queries is to retrieve objects $\{o_1, \ldots, o_n \,|o_i \in \mathcal{X}\}$ that satisfy the proximity conditions relative to the query object q [18]. Among the various types of similarity queries, the k nearest neighbor (k-NN) query is the most popular and widely used [19], [20]. The k-NN query retrieves the k closest objects to a given query object q based on their distance in the metric space. If the dataset, contains fewer than k objects (i.e., $|\mathcal{X}| \leq k$), the query returns the entire dataset (i.e., $|\mathcal{A}| = |\mathcal{X}|$) [17]. The following definition formalizes this query type (see Figure 3).

Definition 4 (k-NN Query). *Let $\mathcal{M} = (\mathcal{S}, d)$ a metric space, $q \in \mathcal{S}$ a query point, and $k \in \mathbb{N}^+$ the required number of answers.*

$$k - \mathcal{NN}(\mathcal{S}, d, q, k) = \{\mathcal{A} \,|\, \mathcal{A} \subseteq \mathcal{X} \subseteq \mathcal{S}, |\mathcal{A}| = k \;\wedge \forall\, (o_i \in \mathcal{A})\, \forall \left(o_j \in \{\mathcal{X} \backslash \mathcal{A}\}\right) : d(q, o_i) \leq d(q, o_j)\}$$

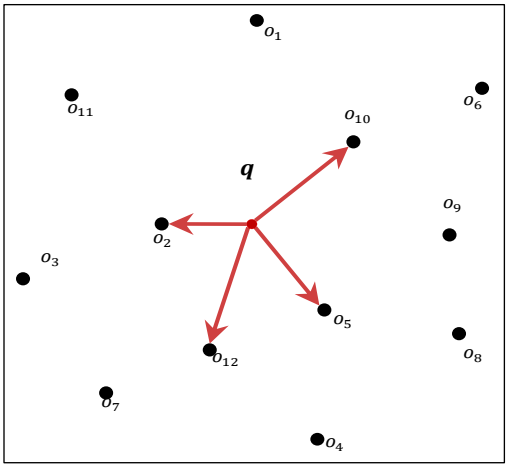


Figure 3: k-NN query, $k - \mathcal{NN}(\mathcal{S}, d, q, k)$.
Source: Authors, (2026).

During the partitioning process, distant points can sometimes be assigned to partitions, leading to a significant increase in their radii r_i. Since these partitions are modeled as hyperballs (see Definition 5), the inclusion of such outlier points can substantially expand the partition's coverage, potentially causing overlap in areas of the data space that do not reflect the actual data distribution.

Definition 5 (Hyperball). *Let $\mathcal{M} = (\mathcal{S}, d)$ a metric space, $\mathcal{X} \subseteq \mathcal{S}$ represents the set of objects. Consider $p \in \mathcal{S}$ as a pivot object, and $r \in \mathbb{R}^+$ as the radius of coverage. The notation $\mathcal{P}(\mathcal{S}, d, p, r)$ defines a closed ball that defines analogously the partitions, where:*

$$\mathcal{P}(\mathcal{S}, d, p, r) = \{\forall o \in \mathcal{X}: d(o, p) \leq r\}$$

Theoretically, overlap between two partitions occurs when the farthest point within a partition (determined by its r_i value) lies farther from the partition's reference point (p_i) than the reference points are from each other. This scenario means that any query ($q \in \mathcal{S}$) within the overlapping space must be processed in each overlapping partition, even if some partitions contain no relevant data at the query location. As a result, distance calculations increase, search response times are prolonged, and computational costs rise, sometimes surpassing those of sequential scans, especially in cases of significant overlap. For example, as illustrated in Figure 4, every partition intersected by the query sphere ($\mathcal{P}_1$ and $\mathcal{P}_2$) must be searched, regardless of whether they contain relevant data points.

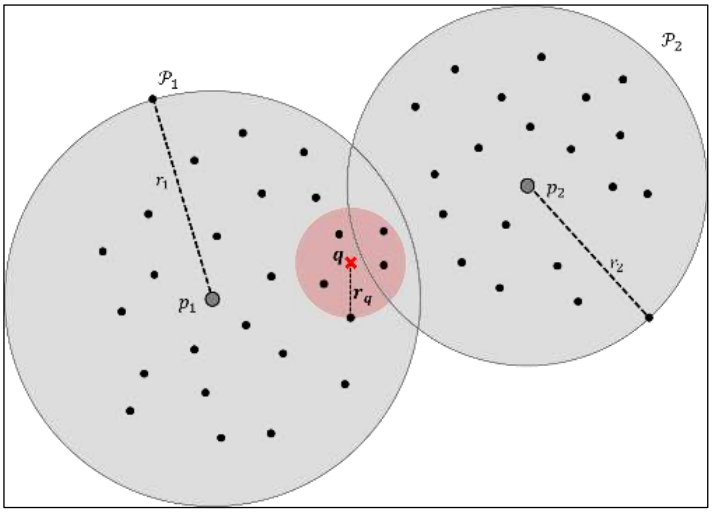


Figure 4: An example of overlapping partitions.
Source: Authors, (2026).

III. RELATED WORKS

In the field of data indexing, numerous techniques have been developed to enhance query efficiency and retrieval speed. These challenges are particularly critical in modern data management systems, especially in IoT. Various indexing techniques have been proposed to address efficient data retrieval in such settings, with several noteworthy structures standing out. One such technique involves Voronoi Diagram-based structures, such as GNAT [21], which leverage Voronoi Diagrams to address node overlaps through static partitioning. While this method effectively prevents node overlaps, it struggles in dynamic environments, particularly when maintaining non-overlapping partitions during real-time updates. The EGNAT [22], a dynamic variant of GNAT and a generalization of the GH-Tree [23], introduces flexible node organization to manage updates following an initial bulk load. However, it continues to face challenges in maintaining the non-overlapping principle of Voronoi diagrams [24], [25], and its search operations can be computationally intensive.

In contrast to static structures, the M-Tree [11] and its variations, such as the M+-Tree [26] and BM+-Tree [27], are known for their balanced structures and effective handling of dynamic updates. Nevertheless, these structures encounter challenges with overlapping nodes at the same level, especially in higher-dimensional spaces. The Slim-Tree [12] builds on the M-Tree by introducing innovative algorithms like the Slim-Down technique to minimize node overlap. However, the effectiveness of overlap reduction varies depending on the characteristics of the data distribution, and the need for object reinsertion during updates adds further complexity. OMNI [28], [29] technology, a notable alternative to the M-Tree, significantly improves search efficiency by using strategically placed vantage points to increase the likelihood of node elimination. However, it still struggles with reducing the number of intersecting nodes, which can negatively impact overall search performance.

Hybrid techniques, such as the IM-Tree [30], XM-Tree [31], and NOBH-Tree [32], combine multiple strategies to mitigate overlap by utilizing pivot-based partitioning and non-overlapping regions. The IM-Tree partitions space around two pivots into five non-overlapping regions, but encounters difficulties when distant regions accumulate too many objects. The XM-Tree addresses this by eliminating partitions without objects, introducing an external limiting area, and using linked lists for storing distant objects. The NOBH-Tree, another hybrid structure, combines Voronoi-shaped and ball-shaped partitions, utilizing two pivots, covering radii, and a hyperplane to create non-overlapping regions. It introduces a sixth region using the hyperplane to manage overlaps caused by the ball-shaped partitions, allowing for flexible tree compositions while ensuring that partitions meet the criteria for non-disjoint regions and symmetric space division relative to the pivots. Despite their potential, these hybrid structures involve complex partitioning and management strategies and do not fully resolve overlap issues, thus limiting their scalability and efficiency in large-scale deployments.

Recently, the BCCF-tree, introduced by [6], presented an innovative approach to real-time IoT data indexing, improving k-NN searches and achieving balanced partitioning using the k-means algorithm. Its extension, the ABCCF-tree, addresses unique labeling challenges in Internet of Video Things (IoVT) networks [33], [34], [35]. Another variant, the QCCF-tree [36], divides space into four regions and enhances query retrieval by enabling concurrent searches at each node, demonstrating superior performance in both construction and similarity searches. However, despite these advancements, these structures still struggle with overlap issues, particularly when managing large-scale datasets.

While the reviewed structures have shown effectiveness in various data environments, they continue to face significant challenges in efficiently managing node overlaps in large databases. Overlap complicates index construction and maintenance, and it undermines query performance by necessitating additional computations and data traversals. These ongoing challenges highlight a clear research gap: the need for a new mechanism that can more effectively quantify, measure, and address the overlap problem. Bridging this gap is essential for improving the efficiency of data indexing techniques.

IV. PROPOSED SOLUTION

Our proposed solution divided into three successive stages. (i) preprocessing, (ii) overlapping estimation, and (iii) decision-making and indexing.

IV.1 PREPROCESSING

To enhance indexing structures, we propose applying a clustering algorithm to the data before indexing. The goal is to organize similar data into compact groups, facilitating efficient indexing and retrieval processes. In our approach, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [37], [38], [39] is selected as the most suitable algorithm, supported by its proven effectiveness in previous studies [13], [38]. DBSCAN identifies clusters of arbitrary shapes based on density, dynamically adjusting without the need to pre-determine the number of clusters, unlike k-means. This method relies on two key parameters: (i) $\epsilon \in \mathbb{R}^+$, the radius of the spherical neighborhood around each object $o \in \mathcal{X}$ (see Definition 6), and (ii) $MinPts$, the minimum number of objects required for a neighborhood to be considered a valid cluster.

Definition 6 (ϵ-Neighborhood of an Object). *Let $\mathcal{M} = (\mathcal{S}, d)$ be a metric space, and $\mathcal{X} \subseteq \mathcal{S}$ represent the set of objects to be grouped. The ϵ-neighborhood of an object o, denoted as $\mathcal{N}_\epsilon(o)$, is defined as:*

$$\mathcal{N}_\epsilon(o) = \{q \in \mathcal{X} \mid d(o,q) \leq \epsilon\}$$

Despite DBSCAN's ability to cluster without requiring centroids or radii for each group, it does not fully align with all our specific needs. Therefore, we will enhance the DBSCAN algorithm (Algorithm 1) to better suit our unique requirements in the subsequent stages of this research.

Algorithm 1: DBSCAN Clustering Algorithm		
Require: $\begin{pmatrix} \mathcal{X} \subseteq \mathcal{S}, \\ d: \mathcal{S} \times \mathcal{S} \to \mathbb{R}^+, \\ \epsilon \in \mathbb{R}^+, \\ MinPts \in \mathbb{N} \end{pmatrix}$ **Ensure:** $\begin{pmatrix} \{\boldsymbol{\mathcal{P}}\} \in \boldsymbol{\mathcal{X}}, \\ \{\boldsymbol{p}\} \in \boldsymbol{\mathcal{S}}, \\ \{\boldsymbol{r}\} \in \mathbb{R}^+ \end{pmatrix}$		
1 :	clusterId ⟵ nextId(noise)	
2 :	**forall** $o \in \mathcal{X}$ **do**	
3 :	**if** o.Cid = UNDEFINED **then**	
4 :	**if** ExpandCluster($\mathcal{X}$, o, clusterId, ϵ, $MinPts$) **then**[1]	
5 :	clusterId ⟵ nextId(clusterId)	
6 :	**endif**	
7 :	**endif**	
8 :	**endfor**	
9 :	$\{\boldsymbol{\mathcal{P}}\}_{i=1}^{clusterId} = \{\forall o \in \mathcal{X} \mid o.Cid = i\}$	
10 :	$\{\boldsymbol{p}\}_{i=1}^{clusterId} = \left\{\frac{1}{\lvert\boldsymbol{\mathcal{P}}_i\rvert}\sum_{o \in \mathcal{P}_i} o\right\}$	
11:	$\{\boldsymbol{r}\}_{i=1}^{clusterid} = \max_{o \in \mathcal{P}_i}\{d(p_i, o)\}$	

Source: Authors, (2026).

IV.2 OVERLAPPING ESTIMATION

In this subsection, we introduce our innovative heuristics designed to minimize the impact of the overlap space problem discussed in Section 2. These heuristics aim to enhance metric space indexing without compromising query retrieval accuracy or the overall quality of the index.

Please note that the function "*ExpandCluster*" from Algorithm 1 remains unchanged. We used it as is and therefore did not include its details here. For further details on "*ExpandCluster*", refer to [37].

IV.2.1 Volume-Based Method (VBM)

The first proposed heuristic is the Volume-Based Method (VBM). This approach provides a detailed measure of overlap by calculating the intersection volume between partitions. As outlined in Section 2, partitions in a metric space are modelled as hyperballs (see Definition (5)). To accurately assess the degree of partition overlap, we measure the overlapping volume between these corresponding hyperballs. The intersection volume between two hyperballs (Figure 5) depends solely on the distance between their centers and the sizes of their radii, as outlined in Definition (7).

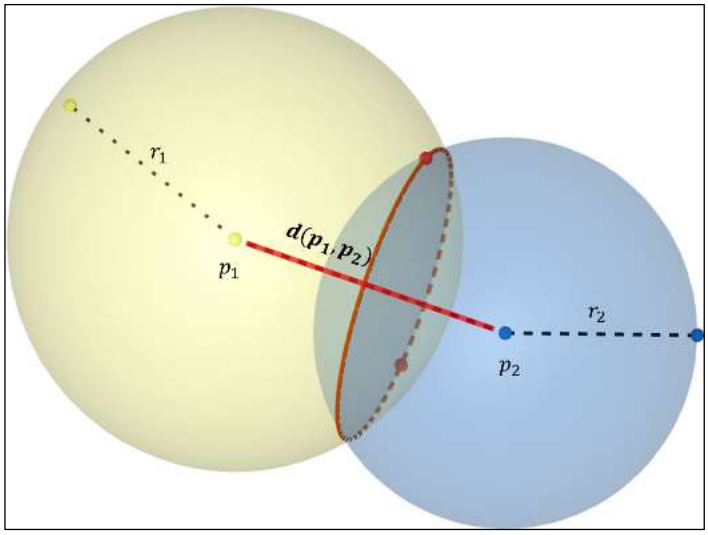


Figure 5: Geometry of the intersection of two hyperballs.
Source: Authors, (2026).

Definition 7. *Let $\mathcal{M}(\mathcal{S}, d)$ a metric space. Consider two partitions $\mathcal{P}_1(\mathcal{S}, d, p_1, r_1)$ and $\mathcal{P}_2(\mathcal{S}, d, p_2, r_2)$, representing hyperballs centered at p_1 and p_2 with radii r_1 and r_2, respectively. The overlap volume (V) and overlap volume rate ($V_\cap$) are defined as*:

$$V = \begin{cases} 0, & if\ d(p_1, p_2) \geq r_1 + r_2 \\ min\{V_{\mathcal{P}_1}, V_{\mathcal{P}_2}\}, & if\ d(p_1, p_2) \leq |r_1 - r_2| \\ \sum_{i=1}^{2} V_{\triangle}(r_i, p_i), & otherwise \end{cases}$$

$$V_\cap = \begin{cases} 0, & if\ d(p_1, p_2) \geq r_1 + r_2 \\ 1, & if\ d(p_1, p_2) \leq |r_1 - r_2| \\ \frac{\sum_{i=1}^{2} V_{\triangle}(r_i, p_i)}{V_{\mathcal{P}_1} + V_{\mathcal{P}_2}}, & otherwise \end{cases}$$

where:

- *$V_{\mathcal{P}_1}$ and $V_{\mathcal{P}_2}$ are the volumes of hyperballs $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively, as defined in Definition (8).*
- *$V_{\triangledown}(r_i, c_i)$ is the volume of the spherical cap of the hyperball $\mathcal{P}_i$ and is calculated using Definition (9).*

Definition 8 (Volume of a Hyperball). *Let $\mathcal{M}(\mathcal{S}, d)$ be a metric space. Consider a partition $\mathcal{P}_i(\mathcal{S}, d, p_i, r_i)$, representing an n-dimensional hyperball centered at c_i with radius r_i. The volume $V_{\mathcal{P}_i}$ is given by [40]:*

$$V_{\mathcal{P}_i} = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2} + 1)} \times r_i^n$$

where Γ denotes the gamma function [41]:

$$\Gamma(x) = \int_0^\infty t^{x-1}\, e^{-t}\, dt$$

As demonstrated in Definition (7), the volume of the intersection between two hyperballs corresponds to the combined volume of the spherical caps formed by a hyperplane cutting through them (see Figure 6).

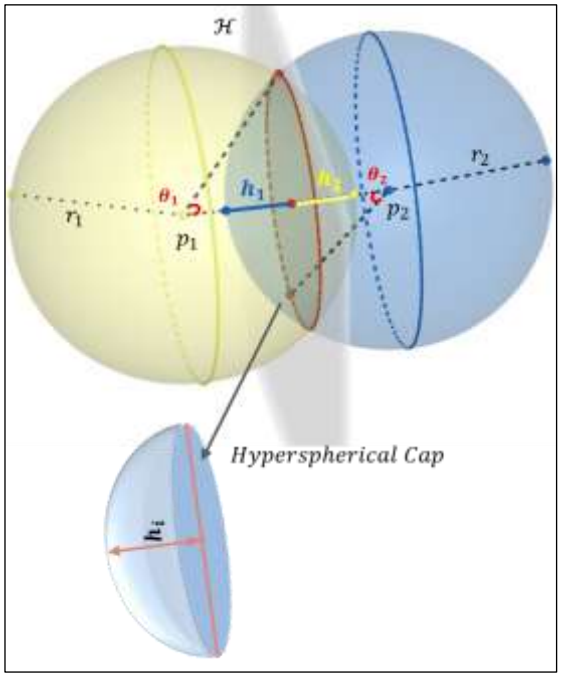


Figure 6: An example of a spherical cap.
Source: Authors, (2026).

To determine the n-dimensional volume $V_{\nabla}(r_i, p_i)$ of a hyperspherical cap with height hi and radius r_i, we apply the methodology described in [40]. The formula for calculating this volume is formalized by the following definition:

Definition 9 (Volume of Hyperspherical Cap). *Let $\mathcal{M}(\mathcal{S}, d)$ denote a metric space. Consider two partitions, $\mathcal{P}_i(\mathcal{S}, d, p_i, r_i)$ and $\mathcal{P}_j(\mathcal{S}, d, p_j, r_j)$, representing n-dimensional hyperballs centered at p_i and p_j with radius r_i and r_j, respectively. The volume of a hyperspherical cap of $\mathcal{P}_i$ with height h_i is calculated using the formula from [40] as follows:*

$$V_{\nabla}(r_i, p_i) = \frac{\pi^{\frac{n-1}{2}} r_i^n}{\Gamma\left(\frac{n+1}{2}\right)} \int_0^{arccos\left(\frac{r_i - h_i}{r_i}\right)} sin^n(\theta_i) d\theta_i$$

where :

- *h_i is the height of the cap, calculated by:*

$$h_i = r_i(1 - cos(\theta_i))$$

- *θ_i is the polar angle between the rays from the center pi of the sphere to the apex of the cap (the pole) and the edge of the disk forming the base of the cap, calculated as:*

$$\theta_i = \cos^{-1} \frac{r_i^2 + d(p_i, p_j)^2 - r_j^2}{2 r_i d(p_i, p_j)}$$

IV.2.2 Distance-Based Method (DBM)

The Distance-Based Method (DBM) is a simpler heuristic for evaluating overlap between partitions by calculating the distance between their centers and radii. Unlike the VBM, the DBM focuses exclusively on distance relationships, ensuring a computationally streamlined process without compromising accuracy. As illustrated in Figure 6, the overlap distance between partitions is represented by the combined height of the caps. Calculating the rate of this overlap, as defined in Definition (10), is crucial for accurately quantifying the degree of overlap. This quantification helps optimize partition boundaries and enhances the efficiency of the indexing structure.

Definition 10 (Overlapping Distance Rate). *Let $\mathcal{M}(\mathcal{S}, d)$ denote a metric space. Consider two partitions, $\mathcal{P}_1(\mathcal{S}, d, p_1, r_1)$ and $\mathcal{P}_2(\mathcal{S}, d, p_2, r_2)$, representing n-dimensional hyperballs centered at p_1 and p_2 with radius r_1 and r_2, respectively. The distance rate of their intersection ($D_{\cap}$) is calculated as follows:*

$$D_{\cap} = \begin{cases} 0, & if\ d(p_1, p_2) \geq r_1 + r_2 \\ 1, & if\ d(p_1, p_2) \leq |r_1 - r_2| \\ \dfrac{\sum_{i=1}^{2} h_i}{d(p_1, p_2)}, & otherwise \end{cases}$$

where h_i represents the height of the cap of hyperball $\mathcal{P}_i$ as defined in Definition (9).

IV.2.3 Object-Based Method (OBM)

The Object-Based Method (OBM) is particularly practical, using the count of objects shared between partitions as a straightforward yet insightful indicator of overlap. This method allows for a more intuitive interpretation of data space dynamics by directly linking overlap to the actual distribution of data. Unlike previous methods that calculate overlapping space—potentially leading to inaccuracies in cases of large but empty overlaps—OBM addresses this issue by focusing on the number of overlapping objects, providing a more accurate representation of data overlap. The mathematical foundation of this approach is illustrated in Definition (11).

Definition 11 (Overlapping Objects Rate). *Let $\mathcal{M}(\mathcal{S}, d)$ be a metric space and let $\mathcal{X} \subseteq \mathcal{S}$ be a set of partitioned objects. Consider two partitions, $\mathcal{P}_1(\mathcal{S}, d, p_1, r_1)$ and $\mathcal{P}_2(\mathcal{S}, d, p_2, r_2)$, representing hyperballs centered at p_1 and p_2 with radius r_1 and r_2, respectively. The set A of objects located in the overlapping area of the two hyperballs is defined by:*

$$A = \mathcal{P}_1(\mathcal{S}, d, p_1, r_2) \cap \mathcal{P}_2(\mathcal{S}, d, p_2, r_2)$$

Explicitly, this is defined as:

$$A = \{o \in \mathcal{X} \mid d(o, p_1) \leq r_1 \; and \; d(o, p_2) \leq r_2\}$$

The rate $A_\cap$ of objects in the overlapping area A is calculated as:

$$A_\cap = \begin{cases} 0, & if\ d(p_1, p_2) \geq r_1 + r_2 \\ 1, & if\ d(p_1, p_2) \leq |r_1 - r_2| \\ \dfrac{|A|}{|\mathcal{P}_1| + |\mathcal{P}_2|}, & otherwise \end{cases}$$

Here,

- *$|A|$ denotes the number of objects in set A.*
- *$|\mathcal{P}_1|$ and $|\mathcal{P}_2|$ represent the number of objects in partitions $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively.*

IV.2.4 Decision-Making and Indexing

After estimating the overlap between partitions, the next step is to create the indexes. We establish two thresholds, and , to quantify the degree of overlap. Based on these thresholds, we categorize overlap into three levels: (i) Low overlap $[0, \xi_{min}]$, (ii) Medium overlap $[\xi_{min}, \xi_{max}]$, and (iii) High overlap $[\xi_{max}, 1]$. These classifications enable us to balance the trade-offs between index complexity and retrieval efficiency by applying tailored strategies to different levels of overlap, thereby ensuring optimal performance. By defining these thresholds, we can adapt our indexing approach to the specific characteristics of the data and the nature of the overlap.

- Low overlap $[0, \xi_{min}]$: In cases of low overlap, we extract all objects from the partition with the smaller cap and merge them into the other partition. This results in two separate indexes, one for each partition. This minimal overlap can be efficiently managed by redistributing objects, which maintains a clear separation between partitions and reduces redundant searches.
- Medium overlap $[\xi_{min}, \xi_{max}]$: For medium overlap, we create three indexes: one for each partition, excluding the overlapping objects, and a third specifically for the overlapping objects. This third index is labeled as a neighbor of the original indexes (partitions) to ensure easy access during the search, particularly when the required number of objects has not yet been reached. This approach provides a balance between merging and maintaining separate partitions. By isolating overlapping objects into a separate index, searches within each primary partition remain efficient while the overlap is accounted for in a dedicated index.
- High overlap $[\xi_{max}, 1]$: In cases of high overlap, we merge the two partitions to create a single index structure for all objects. Significant overlap would lead to excessive redundancy and inefficient searches if the partitions were kept separate. Merging them into a single index simplifies the structure and optimizes retrieval by eliminating the need for extensive overlap management.

The thresholds ξ_{min} and ξ_{max} are crucial for classifying the degree of overlap. Their values directly influence how overlap is quantified and subsequently managed. Setting these thresholds requires careful consideration of the dataset's characteristics and the desired balance between indexing complexity and search efficiency. By finetuning these thresholds, we can develop an optimal indexing strategy that minimizes computational overhead while ensuring accurate and efficient data retrieval. For our index structures, we employ the BCCF-tree [6], an efficient metric index structure designed for big IoT data retrieval. The BCCF-tree partitions the data space into disjoint regions using a recursive k-means algorithm [42], which is effective in minimizing overlap between partitions. However, the construction of the BCCF-tree can be complex due to the recursive nature of the k-means algorithm. By addressing this challenge through the steps outlined earlier, the intricate nature of the k-means algorithm becomes less critical, allowing for the use of a simpler partitioning technique. Therefore, we propose a refined approach utilizing GH partitioning, as defined in Definition 3. GH partitioning is widely recognized for its effectiveness and relatively lower computational complexity compared to other methods. The complexity of hyperplane partitioning typically ranges from $O(n \times log(n))$ to $O(n^2)$, depending on the specific implementation and the dimensionality of the data, where n represents the number of data points [16].

Definition 12 (BCCF-tree). *Let $M(\mathcal{S}, d)$ a metric space, and let $\mathcal{X} \subseteq \mathcal{S}$ represent the set of indexed objects. The BCCF-tree structure is composed of two levels:*

- *Internal nodes $\mathcal{N}$ is a sextuplet:*

$$(p_1, p_2, r_1, r_2, \mathcal{N}_L, \mathcal{N}_R) \in \mathcal{X} \times \mathcal{X} \times \mathbb{R}^+ \times \mathbb{R}^+ \times \mathcal{N} \times \mathcal{N}$$

where:

  - *(p_1, p_2) are two pivots, with $d(p_1, p_2) > 0$.*
  - *(r_1, r_2)represent the distances to the farthest object in the sub-tree rooted at that node $\mathcal{N}$ with respect to p_1 and p_2, respectively, i.e., $r_i = max\{d(p_i, o), \forall o \in \mathcal{N}\}$.*
  - *$(\mathcal{N}_L, \mathcal{N}_R)$ are two sub-indexes (left and right child nodes).*
- *Leaf nodes, denoted as E_C, consist of a subset of the objects stored within a bucket.*

$$E_C \subseteq \mathcal{X}$$
$$where, |E_C| \leq c_{max}, c_{max} = \sqrt{n}, and\ n = |\mathcal{X}|$$

In this paper, we employ the BCCF-tree structure without modification, using the same construction algorithm detailed in [6], [18]. However, during the search phase, we introduce a new algorithm to manage the multiple indexes generated. This algorithm selects the appropriate index or indexes based on their proximity to the query object (see Algorithm 2). Once the closest indexes are identified, the kNN search algorithm of [6], [18] is applied in parallel across all selected indexes, ensuring efficient and accurate query processing while leveraging parallelism for enhanced performance.

Algorithm 2: Recherche kNN

Require: $\begin{pmatrix} indexes \in \mathcal{N}, \\ d: \mathcal{S} \times \mathcal{S} \to \mathbb{R}^+, \\ q \in \mathcal{S}, \\ k \in \mathbb{N}^+, \\ A \in \mathcal{X} = \emptyset \end{pmatrix}$

Ensure: $A \in \mathcal{X}, |A| \leq k$

	*/** **STEP 01:** Estimation of the closest indexes to the query q **/*	
1 :	$min_index \leftarrow 0$	
2 :	$min_dist \leftarrow \infty$	
3 :	$close\ index \leftarrow None$	
4 :	$A \leftarrow \emptyset$	
5 :	**foreach** $index \in indexes$ **do**	
6 :	$dist \leftarrow d(index.center, q)$	
7 :	**if** $dist < min_dist$ **then**	
8 :	$close\ index \leftarrow index$	
9 :	$min_dist \leftarrow dist$	
10 :	**endif**	
11:	**endfor**	
12:	$close_indexes.append(close\ index)$	
13:	**for** $index \in close_index.neighbors$ **do**	
14:	$close\ indexes.append(index)$	
15:	**endfor**	
	*/** **STEP 02:** Run kNN search on the selected indexes **/*	
16:	**forall** $index \in close_indexes$ **do in PARALLEL**	
17:	$r_q \leftarrow Estimated - r_q(index, q, d, r_q = \infty)$	▷ *Run the radius estimate & kNN algorithm of [18]*
18:	$A \leftarrow kNN - BCCF - index(index, q, k, d, r_q, A)$	
19:	$results.append(A)$	
20:	**endforparall**	
	*/** **STEP 03:** Gather results from all parallel processes **/*	
21:	$final_results \leftarrow Gather(results)$	
22:	return $final_results$	

Source: Authors, (2026).

V. SIMULATION AND RESULTS

V.1 OVERLAPPING ESTIMATION

The prototype was implemented in Python on a workstation equipped with an Intel® Core™ i5-4200U CPU at 1.6 GHz, 4 GB of RAM, and running the Linux (Ubuntu) operating system.

Two datasets were used in the experiments. Table 1 provides an overview of the dataset characteristics, along with the parameters used in the DBSCAN algorithm (ϵ and $MinPts$) for each dataset. Additionally, the table includes the fixed overlap thresholds (ξ_{min} and ξ_{max}) used in the experiments.

- DB1: This dataset consists of feature vectors from moving objects obtained from an object tracking simulator in an IoVT environment using wireless cameras [43].
- DB2: Knownasthe Wearable Action Recognition Database (WARD)[2], this dataset is a benchmark database used for recognizing human actions using a wearable motion sensor network [44].

Table 1: Parameter values

Dataset	Size	Dimensions	C_{max}	ϵ	$MinPts$	ξ_{min}	ξ_{max}
Tracking	62,702	20	250	248	250	0.4	0.8
WARD	1,000,000	5	316	91	23		

Source: Authors, (2026).

To comprehensively evaluate the effectiveness of the proposed solution, a series of experiments were conducted, benchmarking its performance against the baseline BCCF-tree structure. The experiments were designed to assess the solution across three key dimensions: (i) Structure evaluation, (ii) Construction cost, and (iii) Search efficiency.

V.2 STRUCTURE EVALUATION

This section provides a detailed analysis of the quality of the indexes generated by the proposed heuristics—DBM, OBM, and VBM—applied to the Tracking and WARD datasets. The analysis primarily evaluates these heuristics based on their effectiveness in managing data space and optimizing tree structures. Key metrics assessed include the distribution of leaf nodes, their balance between the left and right sides of the binary trees, the number of nodes at each index level, and the overall height of the index.

V.2.1 Tracking dataset

In the Tracking dataset, both DBM and OBM performed well in maintaining balanced data distributions within the leaf nodes, with bucket sizes peaking around 250 in most trees and median values ranging from 140 to 215, as illustrated in Figures 7 and 8. However, a slight imbalance was observed in Tree 5, where bucket sizes were significantly smaller. VBM demonstrated its adaptability by effectively managing varying data volumes, with median bucket sizes ranging from 25 to 60 and peaks nearing 110, as seen in Figure 9, underscoring its effectiveness in optimizing tree structures for datasets of different sizes.

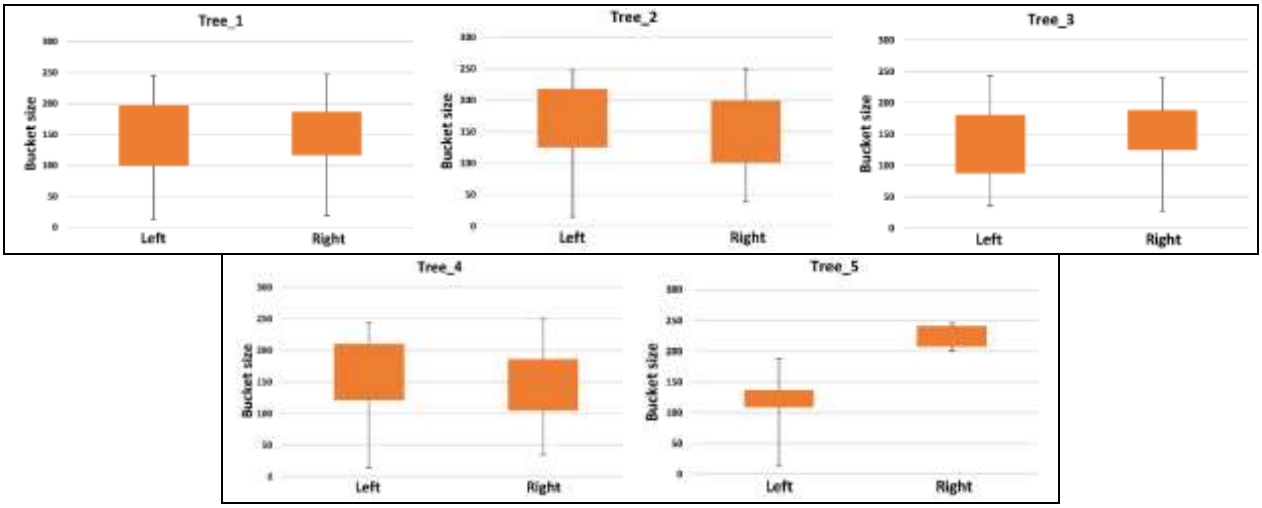
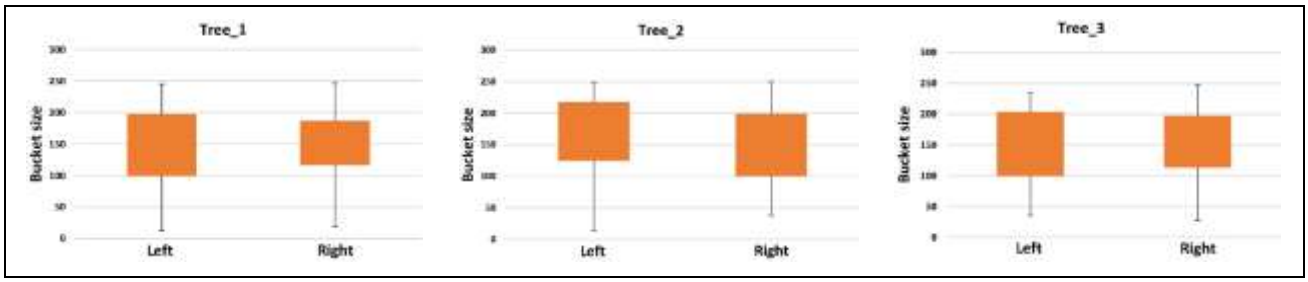


Figure 7: Bucket data distribution using DBM for Tracking datasets.
Source: Authors, (2026).



[2] WARD version 1.0: [download]

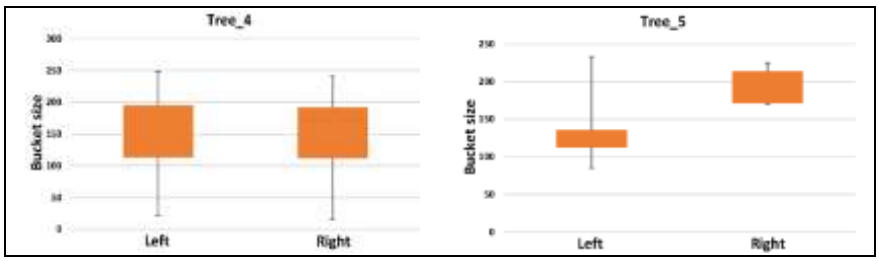


Figure 8: Bucket data distribution using OBM for Tracking datasets.
Source: Authors, (2026).

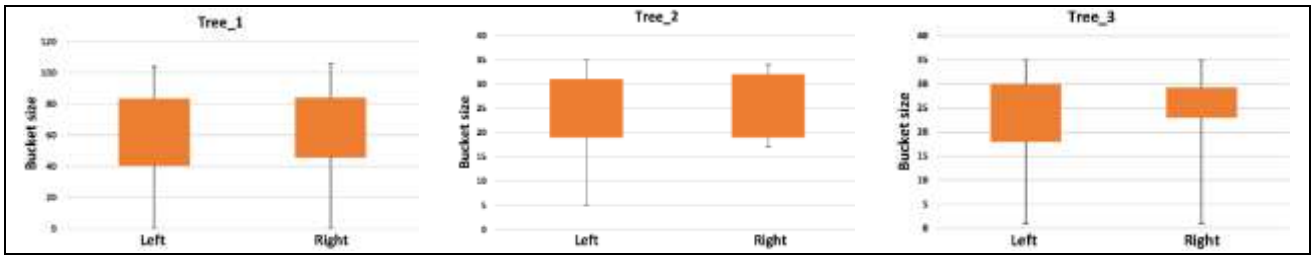
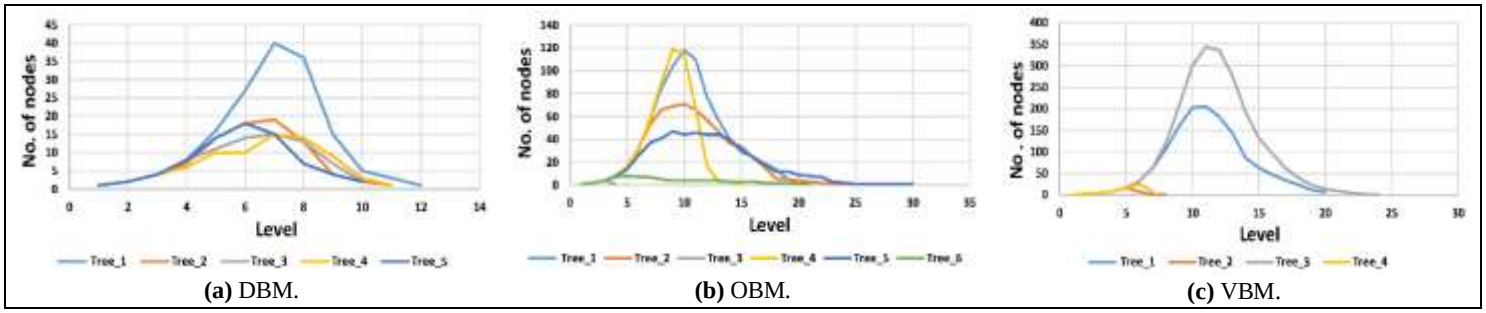


Figure 9: Bucket data distribution using VBM for Tracking datasets.
Source: Authors, (2026).

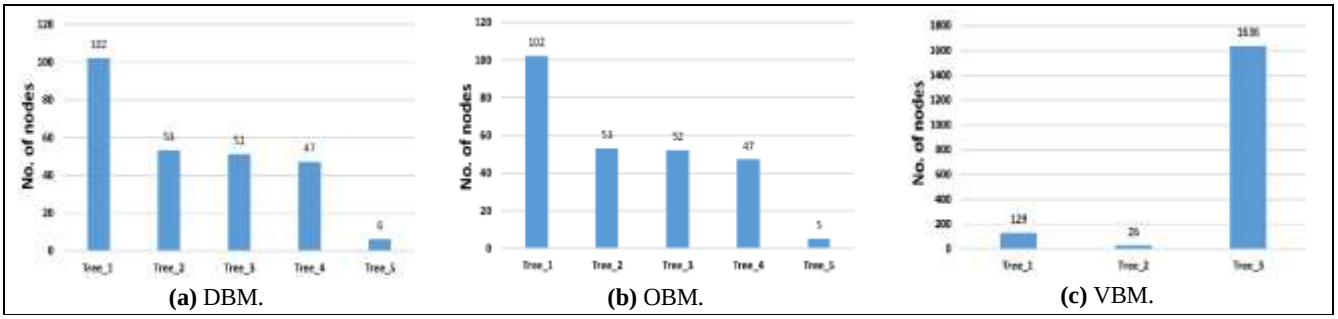
The distribution of nodes at each tree level, shown in Figure 10, further underscores the structural efficiency of these methods. DBM produced trees with a gradual increase in nodes per level, peaking between levels 6 and 8 with a median of 15–20 nodes (Figure 10a). The gradual slope suggests that the tree may remain balanced as the data increases. However, in Tree 5, the slope rapidly decreases, indicating an unbalanced tree, as previously observed. OBM exhibited a similar range of levels, from 6 to 12, but with an increased number of nodes ranging from 17 to 22 and peaks reaching up to 40 nodes. This structure suggests deeper trees with concentrated nodes, which may enhance data partitioning but could also increase traversal costs (Figure 10b). Unlike DBM, OBM's rapid slope suggests that the tree may become unbalanced as the data grows. In contrast, VBM, shown in Figure 10c, displayed a broader node distribution, peaking at around 350 nodes, with levels extending up to 24. The slow slope in VBM indicates its ability to maintain a balanced node distribution across various levels, contributing to a more stable and efficient search process.



(a) DBM. **(b)** OBM. **(c)** VBM.

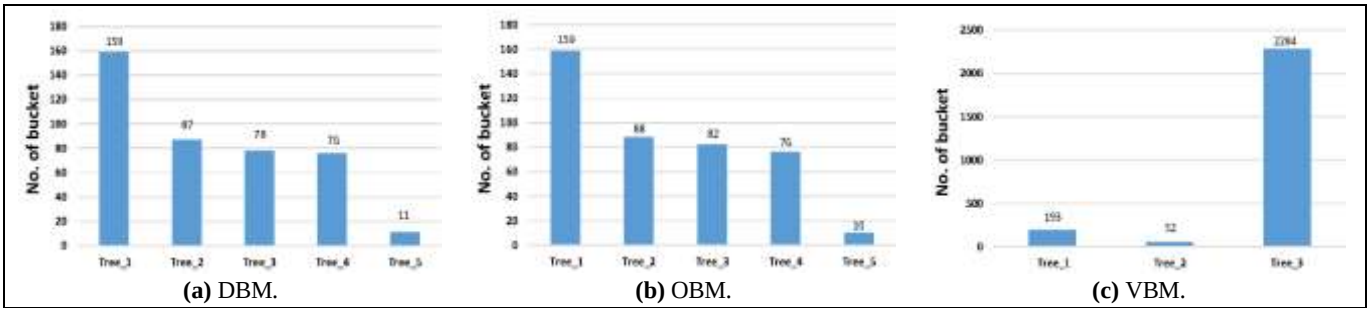
Figure 10: Distribution of nodes across tree levels in the Tracking dataset.
Source: Authors, (2026).

Figures 11, 12, and 13 provide additional insights into the internal structure of the trees. Both DBM and OBM consistently produced balanced trees, with internal nodes ranging from 47 to 102, tree heights between 9 and 12, and leaf nodes ranging from 76 to 159. These results suggest that both methods are effective in maintaining moderate tree depth and efficient partitioning, supporting reliable data retrieval. In contrast, VBM exhibited greater variability, with internal nodes ranging from 26 to 1636, tree heights from 7 to 24, and leaf nodes from 52 to 2284. Despite its larger size across these dimensions, VBM demonstrates a well-filled and balanced structure, making it a highly adaptable solution for managing more complex data distributions.



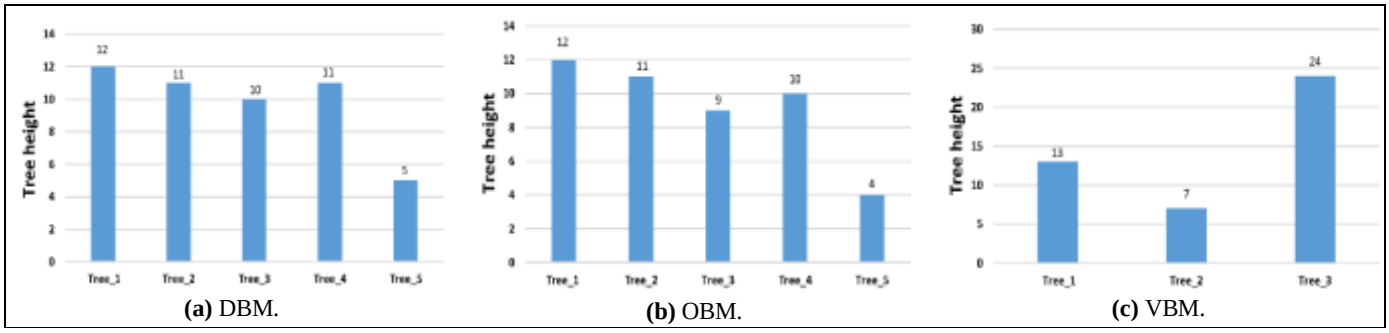
(a) DBM. **(b)** OBM. **(c)** VBM.

Figure 11: Number of Internal Nodes in the Tracking dataset.
Source: Authors, (2026).



(a) DBM. **(b)** OBM. **(c)** VBM.

Figure 12: Number of Nodes Leaves (Buckets) in the Tracking dataset.
Source: Authors, (2026).



(a) DBM. **(b)** OBM. **(c)** VBM.

Figure 13: Height of the Trees in the Tracking dataset.
Source: Authors, (2026).

V.2.2 WARD dataset

In the WARD dataset, all three methods continued to perform strongly, showing consistent results across various metrics. DBM maintained median bucket sizes around 120 to 160 in most trees (Figure 14), with only a slight reduction observed on the left side of Tree 6. OBM exhibited uniform distributions, as seen in Figure 15, with median bucket sizes also around 120 to 160 and a peak around 230. VBM similarly maintained median bucket sizes around 250, with a slight reduction on the left side of Tree 6 (Figure 16).

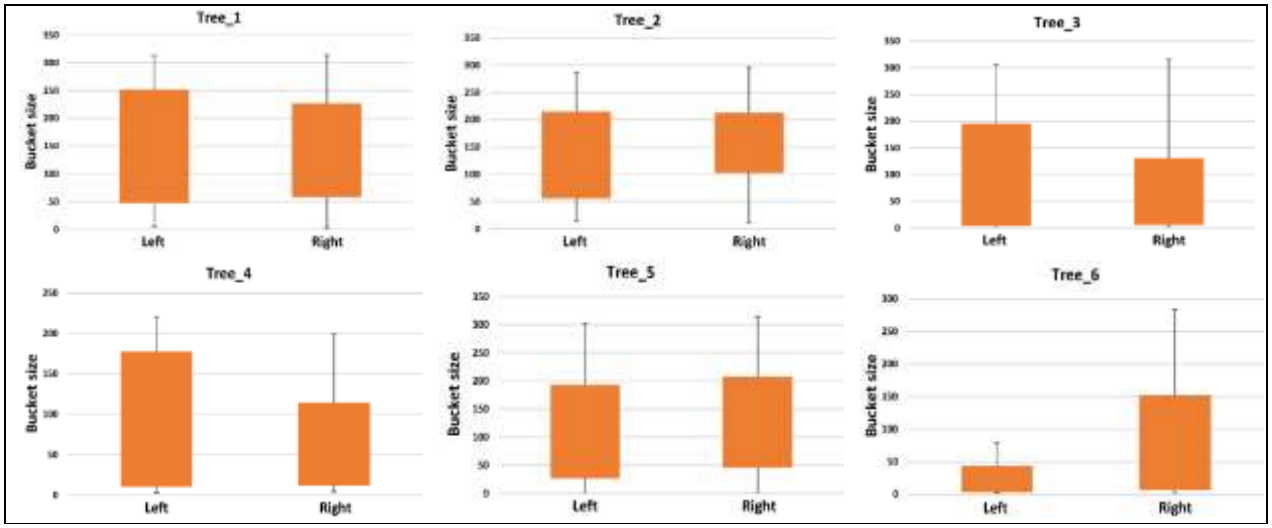
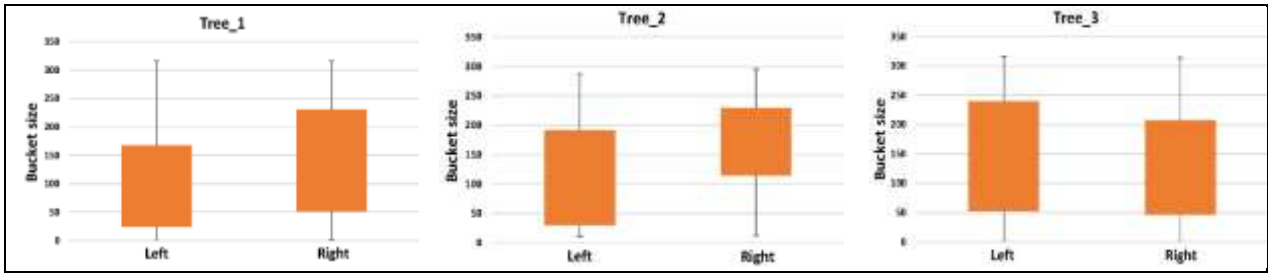


Figure 14: Bucket data distribution using DBM for WARD datasets.
Source: Authors, (2026).



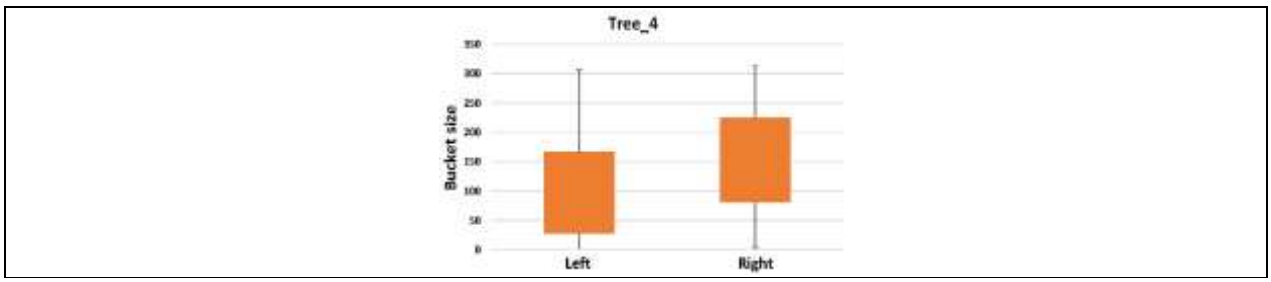


Figure 15: Bucket data distribution using OBM for WARD datasets.
Source: Authors, (2026).

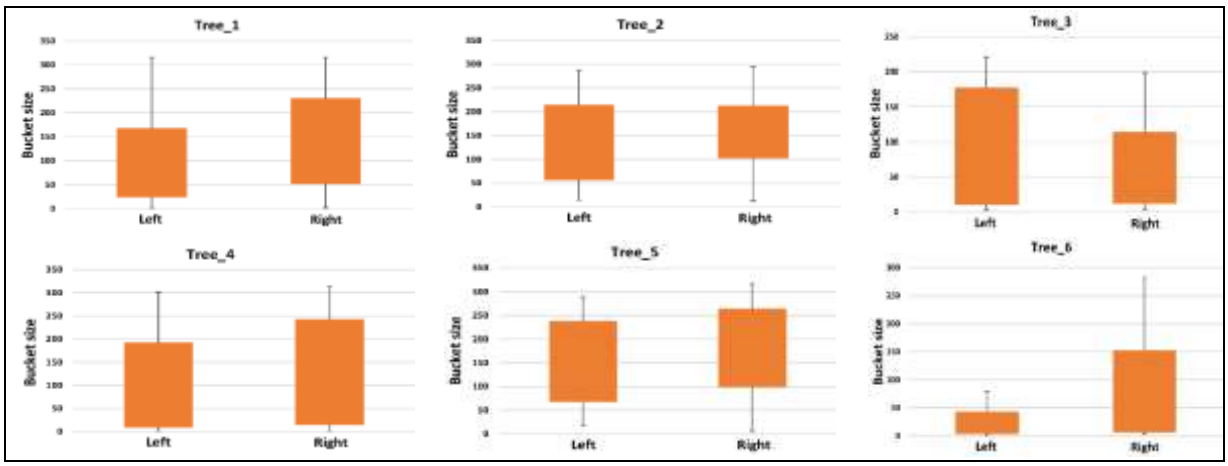
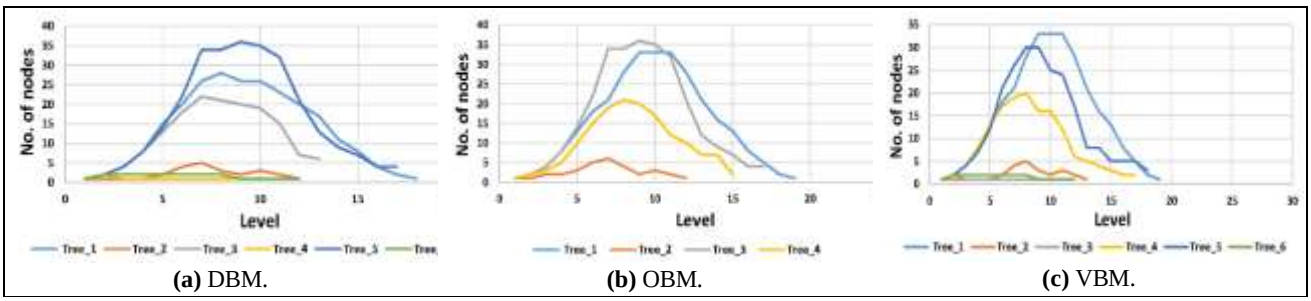


Figure 16: Bucket data distribution using OBM for WARD datasets.
Source: Authors, (2026).

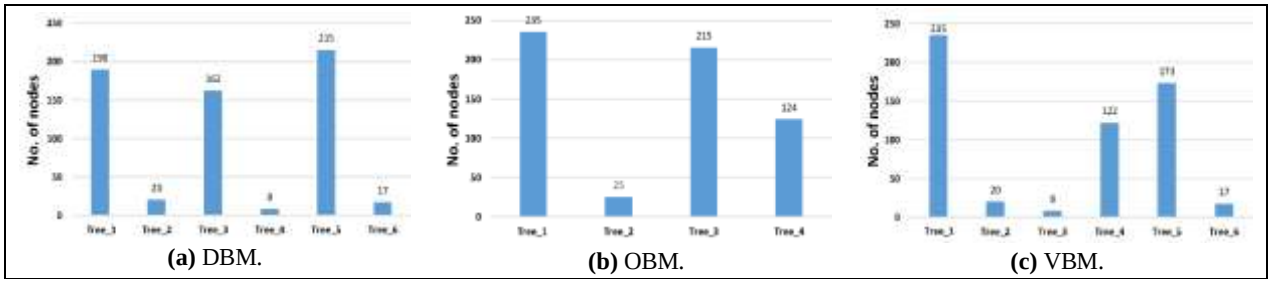
The results of node distribution across tree levels, as shown in Figure 17, reveal consistent performance across all three methods—DBM, OBM, and VBM—with node counts peaking between 30 and 36 at levels 10 to 18. The gradual slopes indicate that each method effectively maintains a balanced distribution of nodes, with tree levels being always well-filled, thereby promoting efficient search operations.



(a) DBM. **(b)** OBM. **(c)** VBM.

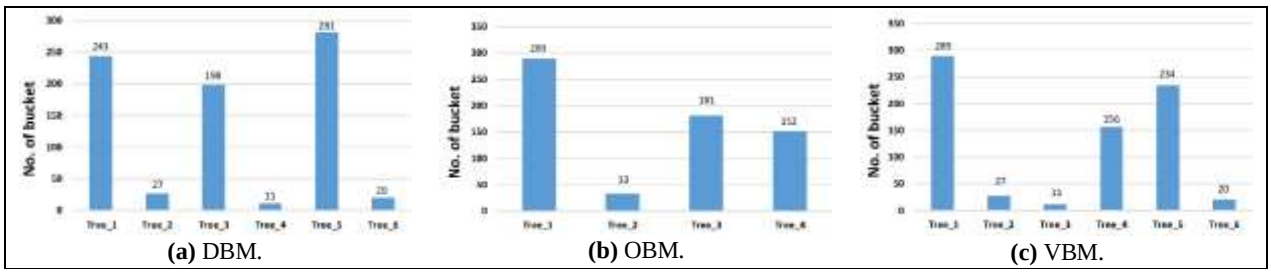
Figure 17: Distribution of nodes across tree levels in the WARD dataset.
Source: Authors, (2026).

Finally, Figures 18, 19, and 20 illustrate the number of internal nodes, leaf nodes, and tree heights for the WARD dataset. DBM and OBM continued to demonstrate balanced structures, with internal nodes ranging from 17 to 215, tree heights between 9 and 19, and leaf nodes ranging from 11 to 281. VBM, once again, displayed a broader range, with internal nodes varying from 8 to 235, tree heights from 9 to 19, and leaf nodes from 11 to 289. This variability highlights VBM's flexibility in managing diverse data complexities and its ability to adapt the tree structure to the specific demands of the dataset.



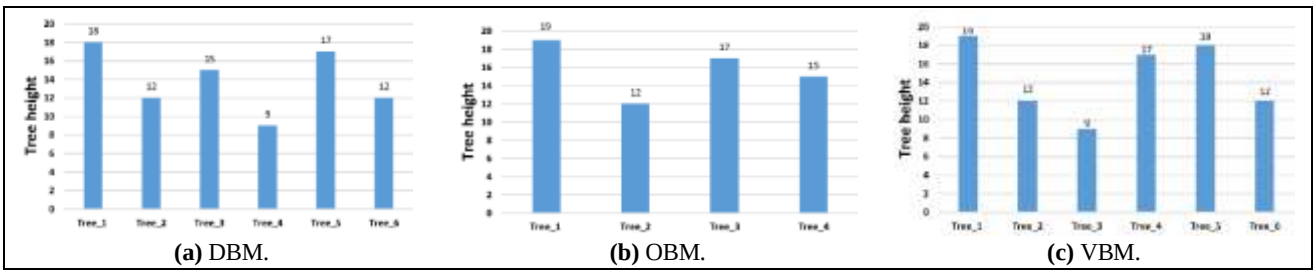
(a) DBM. **(b)** OBM. **(c)** VBM.

Figure 18: Number of Internal Nodes in the WARD dataset.
Source: Authors, (2026).



(a) DBM. **(b)** OBM. **(c)** VBM.

Figure 19: Number of Nodes Leaves (Buckets) in the Tracking dataset.
Source: Authors, (2026).



(a) DBM. **(b)** OBM. **(c)** VBM.

Figure 20: Height of the Trees in the Tracking dataset.
Source: Authors, (2026).

V.3 CONSTRUCTION COST

To evaluate the performance of the proposed structures compared to the BCCF-tree, we analyzed the construction costs in terms of the number of distances calculated and the number of comparisons made during the construction of indexes. The results, presented in Figure 21, provide a clear comparison of these metrics, with the container size fixed as specified in Table 1.

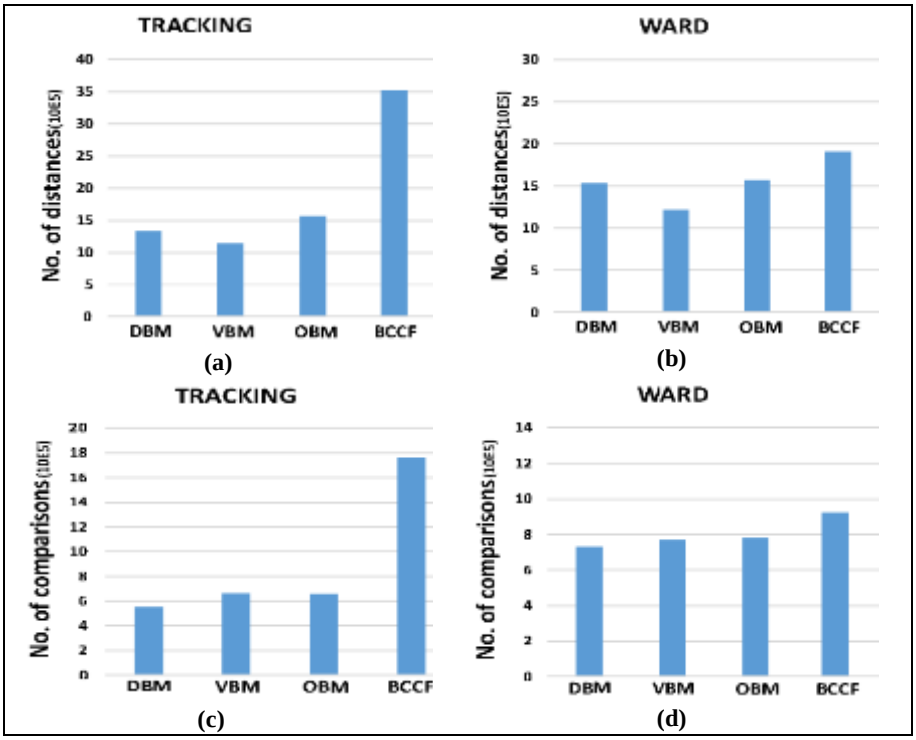


Figure 21: Construction cost.
Source: Authors, (2026).

Across both the Tracking and WARD datasets, the VBM method consistently emerges as the most effective approach. Structural evaluations revealed that VBM produces balanced trees with the flexibility to adapt to varying data complexities, which is directly reflected in its construction costs. In the Tracking dataset, VBM recorded the lowest distance cost at 11.7, compared to 13.6 for DBM and 16.1 for OBM (see Figure 21a), while managing a comparison cost of 6.8, slightly higher than DBM's 5.7 and OBM's 6.5 (see Figure 21c). The trends are reaffirmed in the WARD dataset, where VBM again demonstrates its superiority with a lower distance cost of 12.5, compared to 15.7 for DBM and 16.1 for OBM, as shown in Figure 21b. According to Figure 21d, the comparison cost of VBM in WARD is slightly higher at 7.8, compared to DBM's 7.4, but it remains competitive. Compared to the BCCF-tree, which recorded much higher costs (36.6 in the distance and 17.6 in comparisons), all three methods, particularly VBM, show significant improvements in efficiency across both datasets. This efficiency underscores VBM's ability to maintain a well-distributed structure with minimal overlap.

V.4 SEARCH EFFICIENCY

The similarity search algorithm's performance was evaluated against the BCCF-tree structure by analyzing the average number of distance calculations, comparisons performed, and execution times for fulfilling 100 kNN requests across different k values (k=5, 10, 15, 20, 50, 100). The results of these experiments are presented in Figure 22. In the Tracking dataset, VBM consistently demonstrates superior performance across all metrics. At k = 5, VBM records just 126 distance calculations, significantly lower than DBM's 442 and vastly outperforming OBM's 11,457 and BCCF's 10,573 (see Figure 22a). This trend continues as k increases, with VBM consistently maintaining the lowest distance counts, highlighting its efficiency in managing proximity calculations during searches. In terms of comparisons (see Figure 22c), VBM again outperforms the other methods, recording only 532 comparisons at k = 5, compared to DBM's 1,997, OBM's 13,696, and BCCF's 16,447. This efficiency is further reflected in the search time, where VBM consistently requires the least time across all values of k.

For example, at k = 100, VBM completes the search in just 0.02 seconds, significantly faster than DBM's 0.045 seconds and OBM's 0.07 seconds, while BCCF lags considerably at 1.2 seconds (see Figure 22e). Similarly, in the WARD dataset, VBM continues to outperform the other methods. At k = 5, VBM records 10,686 distance calculations, slightly lower than DBM's 11,268 and significantly better than OBM's 28,717 and BCCF's staggering 191,421 (see Figure 21b). This pattern persists across all k values, with VBM consistently maintaining lower distance and comparison counts, demonstrating its efficiency in processing large datasets. In terms of comparisons, VBM remains strong, with 47,147 comparisons at k = 5, compared to DBM's 50,660, OBM's 129,011, and BCCF's 922,127 (see Figure 22d). This superior performance in both distance and comparison metrics underscores VBM's ability to handle complex data efficiently. Regarding execution time, VBM continues to be competitive, often matching or slightly exceeding DBM's performance while still vastly outperforming OBM and BCCF, especially as k increases (see Figure 22f).

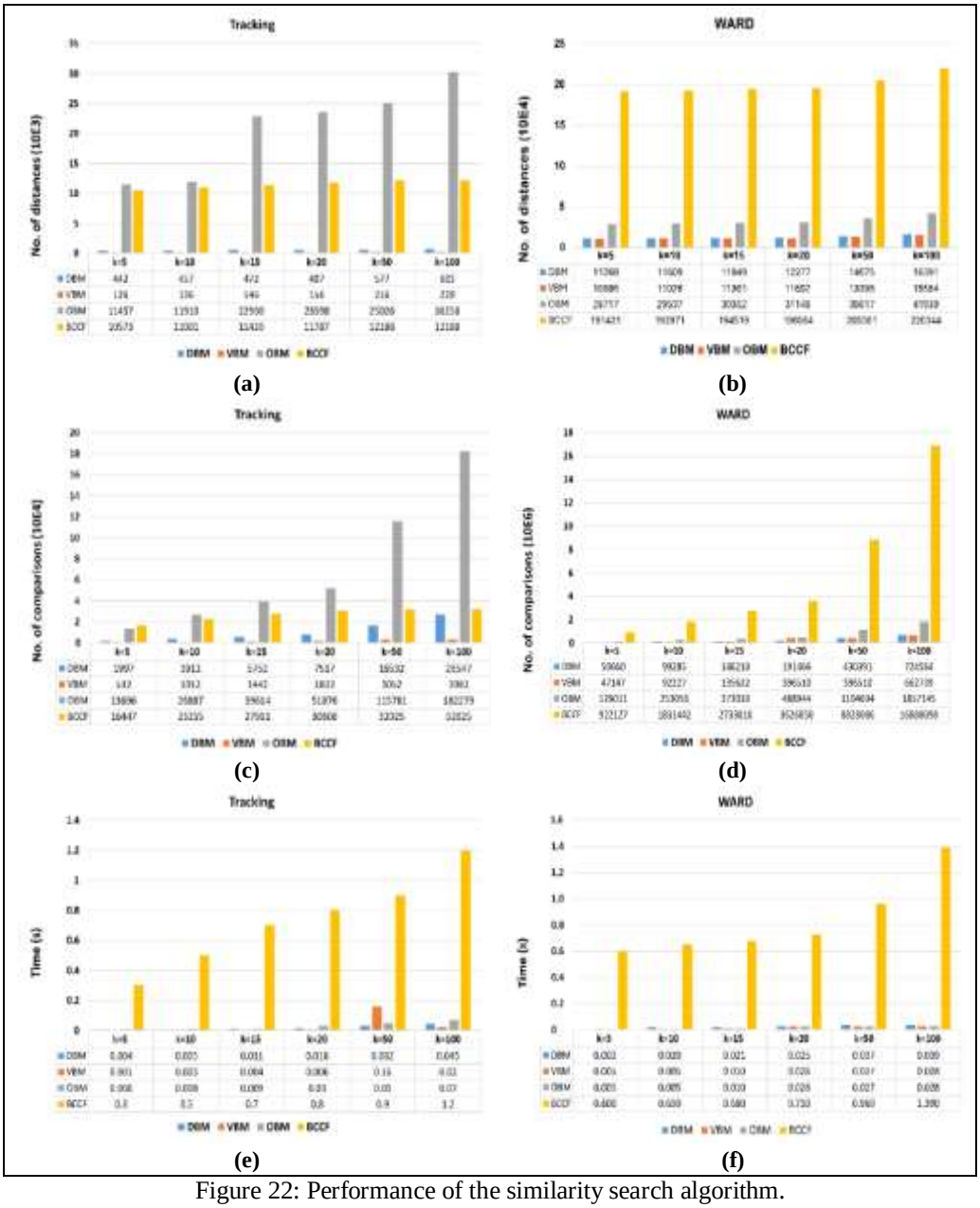


Figure 22: Performance of the similarity search algorithm.
Source: Authors, (2026).

In the final analysis, VBM emerges as the most efficient method for k-NN searches, offering an optimal balance between distance calculations, comparisons, and execution time across both datasets. This superiority is rooted in VBM's advanced approach to overlap management, which minimizes unnecessary computations by precisely calculating intersection volumes, thereby ensuring that search operations remain focused and efficient. DBM, while slightly trailing VBM, demonstrates strong performance with its proximity-based partitioning, which strikes a commendable balance between computational simplicity and effectiveness, though it incurs marginally higher costs. On the other hand, OBM typically incurs increased computational overhead, making it less suitable for large-scale searches. The baseline BCCF-tree consistently underperforms, underscoring the significant advancements and optimizations introduced by the proposed methods, particularly the VBM, in enhancing k-NN search efficiency.

VI. CONCLUSIONS

This paper has addressed the critical challenge of optimizing tree-based indexing structures in the context of massive IoT data. The overlap in data space partitions during index creation has been a significant obstacle, leading to inefficiencies in data retrieval and increased computational costs. To overcome this, we introduced three innovative heuristics: the Volume-Based Method (VBM), the Distance-Based Method (DBM), and the Object-Based Method (OBM), each designed to strategically minimize overlap and enhance the overall performance of the indexing process. Through extensive simulations and comparative analyses, we demonstrated that the VBM consistently outperforms both DBM and OBM, particularly in terms of structural balance, search efficiency, and construction costs. VBM's ability to maintain a well-distributed and adaptable tree structure while minimizing overlap was evident across both the Tracking and WARD datasets, making it the most effective method among the three. DBM followed closely, providing robust performance with slightly higher costs, while OBM, although effective in specific scenarios, generally incurred higher computational overhead due to its detailed partitioning strategy. Moreover, in future work, we aim to further enhance the system's adaptability by exploring dynamic adjustments of overlap thresholds (ξ_{min} and ξ_{max}) tailored to the context and specific characteristics of the data. Additionally, we plan to introduce traceability mechanisms to ensure reliable data monitoring and management across different index structures, optimizing the search space and facilitating more efficient parallel search operations.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Ala-Eddine Benrazek, Ibtissem Kemouguette and Zineddine Kouahla.
Methodology: Ala-Eddine Benrazek and Zineddine Kouahla.
Investigation: Ala-Eddine Benrazek and Brahim Farou.
Discussion of results: Brahim Farou, Zineddine Kouahla.
Writing – Original Draft: Ala-Eddine Benrazek.
Writing – Review and Editing: Brahim Farou and Zineddine Kouahla.
Resources: Ibtissem Kemouguette.
Supervision: Brahim Farou and Hamid Seridi.
Approval of the final text: Ala-Eddine Benrazek, Brahim Farou and Zineddine Kouahla.

VIII. REFERENCES

[1] M. Ge, H. Bangui, and B. Buhnova, "Big data for internet of things: a survey," Future generation computer systems, vol. 87, pp. 601–614, Oct. 2018, doi: 10.1016/j.future.2018.04.053.

[2] K. Ashton, "That 'internet of things' thing," RFID journal, vol. 22, no. 7, pp. 97–114, 2009.

[3] Statista, "Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025," 2016, Statista Hamburg, Germany. Available at: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/ (accessed on Sept. 20, 2025).

[4] R. David, G. John, and J. Rydning, "The digitization of the world from edge to core," Framingham: International Data Corporation, vol. 16, pp. 1–28, 2018.

[5] Y. Liu, M. G. M. Johar, and A. I. Hajamydeen, "IoT-based real time greenhouse monitoring and controlling system," Journal of Engineering and Technology for Industrial Applications, vol. 10, no. 48, pp. 189–195, Jul. 2024, doi: 10.5935/jetia.v10i48.895.

[6] A. E. Benrazek, Z. Kouahla, B. Farou, M. A. Ferrag, H. Seridi, and M. Kurulay, "An efficient indexing for Internet of Things massive data based on cloud-fog computing," Transactions on Emerging Telecommunications Technologies, vol. 31, no. 3, Mar. 2020, doi: 10.1002/ett.3868.

[7] L. C. Shimomura, R. S. Oyamada, M. R. Vieira, and D. S. Kaster, "A survey on graph-based methods for similarity searches in metric spaces," Inf Syst, vol. 95, p. 101507, Jan. 2021, doi: 10.1016/j.is.2020.101507.

[8] L. Chen et al., "Indexing metric spaces for exact similarity search," ACM Comput Surv, vol. 55, no. 6, pp. 1–39, 2022, doi: 10.1145/3534963.

[9] Z. Kouahla et al., "A Survey on Big IoT Data Indexing: Potential Solutions, Recent Advancements, and Open Issues," Future Internet, vol. 14, no. 1, p. 19, Dec. 2021, doi: 10.3390/fi14010019.

[10] D. Jasbick, L. Santos, D. de Oliveira, and M. Bedo, "Some branches may bear rotten fruits: Diversity browsing vp-trees," in International Conference on Similarity Search and Applications, 2020, pp. 140–154. doi: 10.1007/978-3-030-60936-8_11.

[11] P. Ciaccia, M. Patella, P. Zezula, and others, "M-tree: An efficient access method for similarity search in metric spaces," in Vldb, 1997, pp. 426–435.

[12] C. Traina Jr, A. Traina, B. Seeger, and C. Faloutsos, "Slim-trees: High performance metric trees minimizing overlap between nodes," in International Conference on Extending Database Technology, 2000, pp. 51–65. doi: 10.1007/3-540-46439-5_4.

[13] K. Khettabi, Z. Kouahla, B. Farou, H. Seridi, and M. A. Ferrag, “Clustering and parallel indexing of big IoT data in the fog-cloud computing level,” Transactions on Emerging Telecommunications Technologies, vol. 33, no. 7, p. e4484, 2022, doi: https://doi.org/10.1002/ett.4484.

[14] R. Mao, H. Xu, W. Wu, J. Li, Y. Li, and M. Lu, “Overcoming the challenge of variety: big data abstraction, the next evolution of data management for AAL communication systems,” IEEE Communications Magazine, vol. 53, no. 1, pp. 42–47, 2015, doi: 10.1109/MCOM.2015.7010514.

[15] R. Mao, P. Zhang, X. Li, X. Liu, and M. Lu, “Pivot selection for metric-space indexing,” International Journal of Machine Learning and Cybernetics, vol. 7, pp. 311–323, 2016, doi: 10.1007/s13042-016-0504-4.

[16] P. Zezula, G. Amato, V. Dohnal, and M. Batko, “Similarity search: the metric space approach,” vol. 32. Springer Science & Business Media, 2006.

[17] V. Dohnal, “Indexing structures for searching in metric spaces,” Masaryk University, 2004.

[18] A. E. Benrazek, “Internet of Things: Analysis of suspicious behaviour in a surveillance camera network,” University of 08 Mais 1945, Guelma, 2021.

[19] G. Guzun and G. Canahuate, “High-dimensional similarity searches using query driven dynamic quantization and distributed indexing,” Distrib Parallel Databases, vol. 38, no. 2, pp. 255–286, 2020, doi: 10.1007/s10619-019-07266-x.

[20] P. Moutafis, F. Garc\’\ia-Garc\’\ia, G. Mavrommatis, M. Vassilakopoulos, A. Corral, and L. Iribarne, “Algorithms for processing the group K nearest-neighbor query on distributed frameworks,” Distrib Parallel Databases, vol. 39, pp. 733–784, 2021, doi: 10.1007/s10619-020-07317-8.

[21] S. Brin, “Near Neighbor Search in Large Metric Spaces,” in Proceedings of the 21th International Conference on Very Large Data Bases, 1995, pp. 574–584.

[22] G. Navarro and R. Uribe-Paredes, “Fully dynamic metric access methods based on hyperplane partitioning,” Inf Syst, vol. 36, no. 4, pp. 734–747, 2011, doi: 10.1016/j.is.2011.01.002.

[23] J. K. Uhlmann, “Satisfying general proximity/similarity queries with metric trees,” Inf Process Lett, vol. 40, no. 4, pp. 175–179, 1991, doi: 10.1016/0020-0190(91)90074-R.

[24] A. Moriyama, L. S. Rodrigues, L. C. Scabora, M. T. Cazzolato, A. J. M. Traina, and C. Traina Jr, “Vd-tree: how to build an efficient and fit metric access method using voronoi diagrams,” in Proceedings of the 36th Annual ACM Symposium on Applied Computing, 2021, pp. 327–335. doi: 10.1145/3412841.3441915.

[25] M. Marin, R. Uribe, and R. Barrientos, “Searching and updating metric space databases using the parallel EGNAT,” in International Conference on Computational Science, 2007, pp. 229–236. doi: 10.1007/978-3-540-72584-8_30.

[26] X. Zhou, G. Wang, J. X. Yu, and G. Yu, “M+-tree: A new dynamical multidimensional index for metric spaces,” in Proceedings of the 14th Australasian database conference-Volume 17, 2003, pp. 161–168.

[27] X. Zhou, G. Wang, X. Zhou, and G. Yu, “BM+-tree: A hyperplane-based index method for high-dimensional metric spaces,” in Database Systems for Advanced Applications: 10th International Conference, DASFAA 2005, Beijing, China, April
17-20, 2005. Proceedings 10, 2005, pp. 398–409. doi: 10.1007/11408079_36.

[28] A. Traina, C. Traina, C. Faloutsos, and others, “Similarity search without tears: the OMNI-family of all-purpose access methods,” in Proceedings 17th International Conference on Data Engineering, 2001, pp. 623–630. doi: 10.1109/ICDE.2001.914877.

[29] C. Traina, R. F. S. Filho, A. J. M. Traina, M. R. Vieira, and C. Faloutsos, “The omni-family of all-purpose access methods: a simple and effective way to make similarity search more efficient,” The VLDB Journal, vol. 16, pp. 483–505, 2007, doi: 10.1007/s00778-005-0178-0.

[30] Z. Kouahla and J. Martinez, “A new intersection tree for content-based image retrieval,” in 2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI), 2012, pp. 1–6. doi: 10.1109/CBMI.2012.6269793.

[31] Z. Kouahla, A. Anjum, S. Akram, T. Saba, and J. Martinez, “XM-tree: data driven computational model by using metric extended nodes with non-overlapping in high-dimensional metric spaces,” Comput Math Organ Theory, vol. 25, pp. 196–223, 2019, doi: 10.1007/s10588-018-9272-x.

[32] I. R. V. Pola, C. Traina Jr, and A. J. M. Traina, “The NOBH-tree: Improving in-memory metric access methods by using metric hyperplanes with non-overlapping nodes,” Data Knowl Eng, vol. 94, pp. 65–88, 2014, doi: 10.1016/j.datak.2014.09.001.

[33] A. E. Benrazek, Z. Kouahla, B. Farou, H. Seridi, I. Allele, and M. A. Ferrag, “Tree-based indexing technique for efficient and real-time label retrieval in the object tracking system,” J Supercomput, vol. 79, no. 18, pp. 20562–20599, Dec. 2023, doi: 10.1007/s11227-023-05478-8.

[34] I. Allele, A. E. Benrazek, Z. Kouahla, B. Farou, H. Seridi, and M. Kurulay, “Automatic labeling of tracked objects based on an indexing mechanism,” in 2021 International Conference on Theoretical and Applicative Aspects of Computer Science (ICTAACS), 2021, pp. 1–7. doi: 10.1109/ICTAACS53298.2021.9715189.

[35] A. E. Benrazek, Z. Kouahla, B. Farou, H. Seridi, and I. Allele, “Introduction of a tree-based technique for efficient and real-time label retrieval in the object tracking system,” arXiv preprint arXiv:2205.15477, 2022.

[36] K. Khettabi, Z. Kouahla, B. Farou, and H. Seridi, “QCCF-tree: A New Efficient IoT Big Data Indexing Method at the Fog-Cloud Computing Level,” in 2021 IEEE International Smart Cities Conference (ISC2), 2021, pp. 1–7. doi: 10.1109/ISC253183.2021.9562836.

[37] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, and others, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in kdd, vol. 96, no. 34, pp. 226–231, 1996.

[38] I. Kemouguette, Z. Kouahla, A.-E. Benrazek, B. Farou, and H. Seridi, “Cost-Effective Space Partitioning Approach for IoT Data Indexing and Retrieval,” in 2021 International Conference on Networking and Advanced Systems (ICNAS), IEEE, Oct. 2021, pp. 1–6. doi: 10.1109/ICNAS53565.2021.9628904.

[39] S. O. Al-mamory and I. S. Kamil, “A new density based sampling to enhance dbscan clustering algorithm,” Malaysian Journal of Computer Science, vol. 32, no. 4, pp. 315–327, Oct. 2019, doi: 10.22452/mjcs.vol32no4.5.

[40] S. Li, “Concise formulas for the area and volume of a hyperspherical cap,” Asian Journal of Mathematics & Statistics, vol. 4, no. 1, pp. 66–70, 2010, doi: 10.3923/ajms.2011.66.70.

[41] E. Artin, The gamma function. Courier Dover Publications, 2015.

[42] E. L. Lydia et al., “Clustering and indexing of multiple documents using feature extraction through apache hadoop on big data,” Malaysian Journal of Computer Science, vol. 2020, no. Special Issue 1, pp. 108–123, Nov. 2020, doi: 10.22452/mjcs.sp2020no1.8.

[43] A. E. Benrazek, B. Farou, Z. Kouahla, M. A. Ferrag, and H. Seridi, “IoVT-based efficient solution for optimal active smart camera selection in a tracking mission,” Internet of Things, vol. 24, p. 100941, Dec. 2023, doi: 10.1016/j.iot.2023.100941.

[44] A. Y. Yang, R. Jafari, S. S. Sastry, and R. Bajcsy, “Distributed recognition of human actions using wearable motion sensor networks,” J Ambient Intell Smart Environ, vol. 1, no. 2, pp. 103–115, 2009, doi: 10.3233/AIS-2009-0016.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 104-114. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2753

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

MACHINE LEARNING FOR CORONARY HEART DISEASE RISK PREDICTION: A CONVENTIONAL AND ENSEMBLE CLASSIFICATION MODEL APPROACH

Ch V Raghavendran*[1], Kathi Chandra Mouli[2], S. Bhargavi Latha[3]

[1]Department of IT, Aditya University, Surampalem, AP, India.
[2]Department of CSE, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Bachupally, 500090, India.
[3]Associate Professor, School of CSE, Reva University, Yelahanka, Karnataka 560064, India.

[1]https://orcid.org/0000-0002-4423-9086, [2]https://orcid.org/0009-0001-0889-9263, [3]https://orcid.org/0000-0002-0102-3704

Email: *raghuchv@yahoo.com, kcm7961.1852@gmail.com, s.bhargavilatha@gmail.com

ARTICLE INFO

Article History
Received: September 28, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Coronary Heart Disease,
Classification,
Ensemble techniques,
Class imbalance,
Boosting,
Bagging,
Voting,
Stacking.

ABSTRACT

Coronary Heart Disease (CHD) is the leading cause of illness and mortality worldwide. This is driven by risk factors such as high cholesterol, hypertension, smoking, diabetes, obesity and genetic predisposition. Early detection and management of CHD are critical and involve lifestyle changes, pharmacological interventions, angioplasty, or bypass grafting. Traditional diagnostic methods require effective and timely medical interventions or the early detection of CHD. The proposed work suggests a stacking classifier for early prediction of CHD. Conventional classification models are typically simple, interpretable and efficient, making them suitable for classification problems. The performance of these models is limited when dealing with complex or high-dimensional data. Ensemble models address this limitation by combining the predictions of several models to improve accuracy. Bagging, Voting, and stacking classifiers are more sophisticated ensemble methods that use this concept further by using a metamodel to aggregate the predictions of multiple base models. In this study, several conventional models, such as the SGD Classifier, Logistic Regression, K-Nearest Neighbor, Naive Bayes, Decision Tree, and Random Forest Classifiers are applied to the dataset to predict whether the patient has 10 year risk of Coronary Heart Disease (CHD). Boosting models such as Gradient Boosting, AdaBoost, and XGBoost are also applied. The proposed voting classifier resulted in better accuracy, precision, recall, and other metrics compared with the conventional, boosting, and bagging models.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Coronary Heart Disease (CHD), also known as ischemic heart disease, is one of the most common cardiovascular diseases and a major cause of death worldwide. Atherosclerosis, a disorder marked by the build-up of fatty deposits, cholesterol, and other materials in the coronary arteries, is the main cause of CHD, which is a condition that impairs blood flow to the heart muscle. The myocardium receives less oxygen when these arteries gradually narrow or get blocked, it may cause myocardial infarction, angina pectoris, and sudden cardiac death, among other clinical signs. [1]. With millions of deaths annually, CHD continues to be a major global public health concern. The World Health Organization (WHO) has reported that 32% of fatalities worldwide are attributable to cardiovascular illnesses, including CHD. Low- and middle-income countries have disproportionately high illness burdens because they have less access to healthcare and preventive measures. Globally, the incidence of CHD is on the rise owing to aging populations and the rising prevalence of risk factors including diabetes and obesity. The persistent inflammatory condition of atherosclerosis inside arterial walls functions as the main cause of CHD. Chemicals from LDL cholesterol enter the endothelium to form deposits inside artery walls during the first stage because of endothelial dysfunction. An immunological reaction gets launched after oxidized LDL draws monocytes and macrophages which later turn into foam cells that help to form fatty streaks [2]. Artery lumen becomes narrower because fibrous plaques form out of previously developed streaks and reduce blood flow. The formation of a thrombus after plaque rupture leads to total artery blockage thus

causing heart attack or other acute coronary syndromes [3]. The development of CHD depends on both risk factors that people can change and those that cannot be changed. Conflicting variables that people can change include smoking alongside high blood pressure and high cholesterol together with diabetes and obesity while a sedentary lifestyle and poor eating choices complete the list [4]. Two important non-controllable CHD risk factors involve sex-based differences (men face greater risk) and the existence of cardiovascular disease in family history. The development of CHD depends on several newly emerging risk variables that include stress factors together with persistent inflammatory responses and particular gene sequences. The risk of CHD can be considerably decreased by effectively managing modifiable risk factors. Asymptomatic phases to severe, life-threatening illnesses are only a few of the many clinical presentations of congenital heart diseases. Palpitations, exhaustion, shortness of breath, and chest pain or discomfort (angina) are typical symptoms [5]. Stable angina is typified by consistent chest discomfort caused by emotional or physical stress, whereas unstable angina occurs suddenly and is related to a greater risk of myocardial infarction. Prolonged chest discomfort, nausea, sweating, and, in extreme situations, unconsciousness are typical symptoms of myocardial infarction, sometimes known as a heart attack.

Imaging methods, laboratory testing, and clinical evaluation are all used for the diagnosis of CHD. While biomarkers, such as troponin levels, aid in confirming myocardial injury, electrocardiography (ECG) and stress tests are frequently used to evaluate heart function. Imaging techniques that show the amount of blockage and provide thorough visualization of the coronary arteries include echocardiography, computed tomography (CT) angiography, and coronary angiography. Non-invasive techniques for risk assessment and early subclinical disease diagnosis are provided by emerging technologies, such as coronary artery calcium scoring. Medical treatment of CHD requires surgical interventions combined with drugs along with changes to individual lifestyles. Sustaining healthy changes in lifestyle including stopping smoking while starting physical exercise and maintaining proper weight and eating well remain fundamental for slowing disease advancement. Patients with CHD need beta-blockers and statins in combination with ACE inhibitors as well as antiplatelet medications to manage their symptoms while reducing health risks [6]. The medical procedure of both CABG and PCI serves to rebuild cardiovascular blood flow in conditions where substantial artery obstructions occur [7]. Current developments in personalized medicine united with risk-prediction tools have boosted the accuracy of managing CHD.

Primary prevention strategies aim at managing risk factors throughout the general population to reduce CHD incidence [8]. The primary prevention approach relies on three foundational elements that consist of prompt hypertension and hyperlipidemia testing along with community-wide healthy life education and expanded preventive health care facilities. The prevention of future cardiovascular events requires secondary prevention protocol to address patients with high-risk CHD profiles or established CHD through intense risk-factor management and scheduled follow-up. In eight villages in Indonesia's East Java Province's Malang District, Maharani A et al. [9] performed comprehensive household screenings for cardiovascular threat issues in persons aged 40 years and older. The World Health Organization/ International Society of Hypertension's region-specific charts, which take into account factors such age, sex, blood pressure, diabetes status, and smoking behavior, were used to create 10-year cardiovascular risk scores. Adults in Indonesia who are 40 years or older are at high cardiovascular risk, and preventative treatment rates are poor. Population-based and clinical strategies for CVD prevention should be emphasized in both urban and rural regions. This work applies various classification algorithms and ensemble models for the early diagnosis of CHD based on the number of features including sex, age, smoking habit, diabetes, blood pressure, heart rate, and glucose. Section II summarizes the existing work in the field of CHD and the use of conventional and ensemble classification algorithms. Section III discusses the proposed model and the results of the data preprocessing techniques applied to the dataset. The experimental investigations and results are discussed in Section IV. Conclusions of the work are stated in Section V.

II. RELATED WORK

For Heart Disease Prediction (HDP), Jawalkar, A.P., et. al., [10] suggested a method using a Decision Tree-based Random Forest (DTRF) classifier with loss optimization. The findings show that the proposed HDP-DTRF method outperforms conventional techniques, achieving an 85% F1-score, and 96% accuracy on publicly accessible datasets. H. Yang et.al, [11] proposed the HY_OptGBM prediction model, which uses the optimized LightGBM classifier to predict CHD. The most sophisticated hyperparameter optimization framework (OPTUNA) was used in this study to optimize the hyperparameters of the prediction model. CHD data from the Framingham Heart Institute's CHD data was used in this study to assess the prediction model. The proposed model outperformed other comparative models with an AUC score of 97.8%. C. Bemando et. al. [12] developed coronar, the model draws its information from public Cleveland heart disease patient data located at UCI repository. The study confirms that Random Forest achieves 75.00% accuracy while Gaussian Naïve Bayes and Bernoulli Naïve Bayes deliver 85.00% accuracy each. The data analysis demonstrates that the Gaussian and Bernoulli Naïve Bayes models show superior performance compared to Random Forest regarding precision and recall and F-measure which proves their worth for treatment forecasting. The research by Dulani et al. [13] provides an extensive system design approach which combines machine learning methods with Explainable AI (XAI) for developing dependable Coronary Artery Disease (CAD) prediction models.

The main purpose combines two critical tasks: prediction model development in conjunction with SHapley Additive exPlanations (SHAP) analysis which expands model understanding. Different machine learning algorithms were applied through careful system design to determine which model is best for CAD prediction. This study emphasizes the significance of interpretability and predicted accuracy in medical decision making. Shah, D., et. al. [14] discussed heart disease-related features and a model based on algorithms such as random forest, Naïve Bayes, decision tree, and K-nearest neighbor algorithms. This makes advantage of the Cleveland database, which is a UCI archive of people with cardiac illness with 303 samples and 76 features. Only 14 of these 76 characteristics were considered for testing, which is crucial for verifying the effectiveness of various algorithms. The findings showed that K-nearest neighbor yielded the highest accuracy score. In order to provide supplemental medical services, Shan Xu et. al. [15] concentrated on developing a more precise and useful risk prediction system. Both the Cleveland Heart-Disease Database (CHDD) and the PKU People's Hospital Cardiology inpatient dataset were examined to verify their accuracy and usefulness. Our system's accuracy of 91.6% in the CHDD test is noticeably greater than that of other approaches. With the exception of SVM (98.9%), it outperforms most other classifiers with an accuracy of 97% in the People's Hospital dataset test; nevertheless, random forest takes half as long as SVM. A. N. Repaka et.al., [16] focused on diagnosing cardiac disease by taking into account prior data and knowledge to build Smart Heart Disease Prediction (SHDP) using Navies Bayesian

that predicts heart disease risk variables. Dataset collection, application-based user registration and login, Navies Bayesian classification, prediction, and secure data transfer using AES (Advanced Encryption Standard) are the steps in the suggested methodology. The results show that the current diagnostic approach predicts risk factors for cardiac illnesses in an effective manner. Zhenzhen Du et.al., [17] sought to develop a high-precision CHD forecast method using big data and machine learning techniques, based on a sizable group of hypertension patients in Shenzhen, China. With an area under the receiver operating characteristic curve (AUC) score of 0.943, the ensemble approach XGBoost demonstrated significant accuracy in predicting the beginning of CHD within three years for the independent test dataset. On the same datasets, comparison study revealed that machine-learning techniques greatly outperformed fixed models or traditional risk scales, while nonlinear models (K-nearest neighbor with AUC 0.908, random forest with AUC 0.938) perform better than linear models (logistic regression with AUC 0.865). The development of an accurate predictive model by Bhatt CM et.al, [18] aims to reduce cardiovascular disease death numbers. The accuracy rates can rise through implementation of k-modes clustering with Huang beginning as a suggested technique. The analysis utilizes XGBoost (XGB) and Multilayer Perceptron's (MP) as well as Random Forest (RF) and Decision Tree classifier (DT). The evaluation of Area Under the Curve show 0.94 for decision trees alongside 0.95 for XGBoost and 0.95 for random forests and 0.95 for multilayer perceptron's. The accuracy analysis of multilayer perceptron with cross-validation has demonstrated the superior performance among all employed methods reaching 87.28% accuracy.

A combination of a pre-trained Deep Neural Network (DNN) followed by Principal Component Analysis (PCA) for dimensionality reduction through which the prediction occurs by a Logistic Regression (LR) model was presented by Diman Hassan et.al., [19] as a novel approach for Heart Disease (HD) prediction. A DNN + PCA + LR combination was tested on Cleveland HD dataset which is available to the public. The proposed methodology achieved superior accuracy rates compared to existing methods through most measurements according to evaluation results. Mert Ozcan and Serhat Peker analyzed the Classification and Regression Tree (CART) algorithm that serves as a supervised machine learning method for finding decision rules which connect input data to output values and performs cardiac disease predictions [20]. Furthermore, the study's conclusions prioritize the characteristics that affect heart disease. The model's dependability is confirmed by the prediction's 87% accuracy when all performance metrics are taken into account. On testing data, which makes up 85.70% of the entire NHANES dataset, Aniruddha Dutta et.al., [21] developed a CNN architecture with a classification power of 77% to properly classify the existence of CHD patients and 81.8% to successfully classify the nonappearance of CHD cases. This outcome shows that other healthcare studies with a comparable sequence of features and imbalances can use the suggested architecture. Although our suggested CNN model's recall values are similar to those of other machine learning techniques like SVM and random forest, our model is more accurate at predicting negative (non-CHD) cases. C. Beulah Christalin Latha and S. Carolin Jeeva [22] studied ensemble classification, a technique that combines several classifiers to increase the accuracy of weak algorithms. The study's findings show that ensemble methods, such bagging and boosting, perform satisfactorily in determining heart disease risk and are useful in increasing the prediction accuracy of weak classifiers. Ensemble classification helped weak classifiers achieve an accuracy gain of up to 7%.

III. DATA PRE-PROCESSING AND PROPOSED MODEL

The machine learning pipeline requires data pre-processing as its essential first step in order to increase algorithm performance while promoting data quality standards. Data pre-processing requires the transformation of unorganized data into ready-to-train format which organizes information into a structured shape. Regular data pre-processing duties involve encoding categorical data into numbers together with outlier adjustment and feature scaling to normalize ranges and different methods to handle missing data values [23] [24]. Multiple techniques such as feature selection and data augmentation and dimensionality reduction help both boost model effectiveness and stop overfitting phenomena. The pre-processing of data enhances machine learning algorithm efficiency and prediction accuracy because it helps find meaningful patterns [25]. This research delivers a system to detect early coronary heart disease onset among patients and non-patients. Multiple ensemble techniques enhanced both accuracy and prediction stability throughout refined classification modelling processes on the Otto dataset. Bagging, voting, and stacking techniques were used to build ensemble frameworks, and the outcomes of these techniques were carefully examined. The process flow of the suggested model is shown in Figure 1.

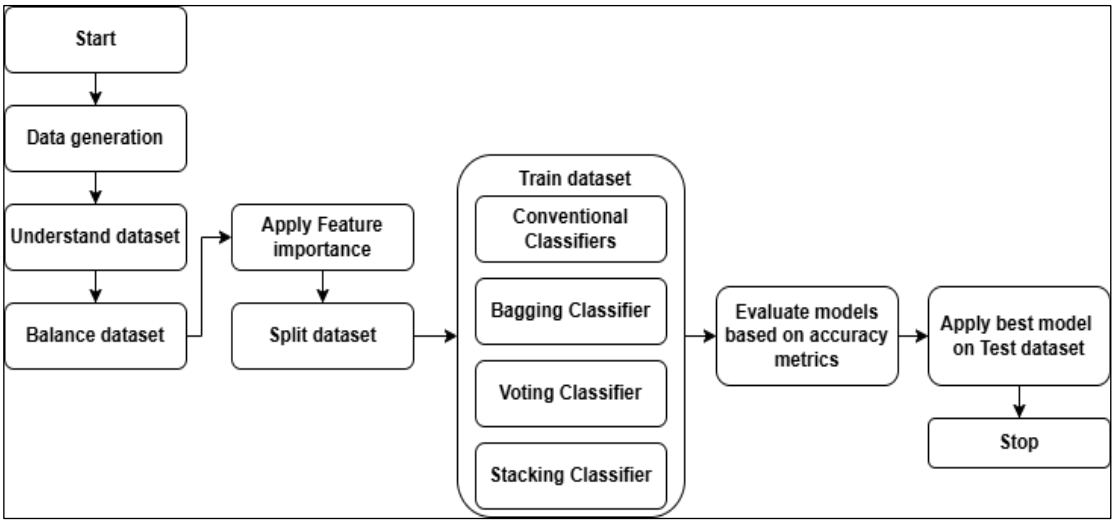


Figure 1: Steps in the proposed model.
Source: Authors, (2026).

III.1 DATASET AND ITS FEATURES

The "Framingham" dataset, which comes from a current cardiovascular study on the people of Framingham, Massachusetts, is openly accessible on the Kaggle website [26]. Determining the patient's 10-year risk of increasing coronary heart disease is the classification goal. Figure 2 displays details about the five patients in a dataset consisting of 4,240 records and 15 features.

male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0

Figure 2: Sample dataset.
Source: Authors, (2026).

III.2 PERFORMANCE METRICS

Classification algorithm performance measures are used to assess the performance of the model with respect to predicting the target classes. The most popular metrics are as follows:

Accuracy – In this measure the ratio of successful instance classifications reflects all total instances.

$$Accuracy = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{Total\ Instances}$$

Precision – The proportion of correctly predicted positive instances among all instances predicted as positive.

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}$$

Recall (Sensitivity or True Positive Rate) – The proportion of correctly predicted positive instances among all actual positive instances.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

F1-Score – The harmonic mean of precision and recall, providing a balanced metric.

$$\text{F1} - \text{Score} = 2\text{x}\frac{\text{Precision x Recall}}{\text{Precision} + \text{ Recall}}$$

Specificity (True Negative Rate) – The rate at which negative cases are predicted correctly out of all the real cases of negative.

$$\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)} + \text{False Positives (FP)}}$$

ROC-AUC (Receiver Operating Characteristic – Area Under the Curve) – This measurement evaluates how effectively the model manages true positive recognition rates (Recall) and wrong positive classifications. A model achieves higher performance when it displays increased AUC values.

Log Loss (Logarithmic Loss) – Quantifies the uncertainty of the model's predictions. Lower values indicate better performance.

$$\text{Log Loss} = -\frac{1}{\text{N}}\sum_{\text{i=1}}^{\text{N}}[\text{y}_\text{i}\log(\text{p}_\text{i}) + (1 - \text{y}_\text{i})\log(1 - \text{p}_\text{i})]$$

Confusion Matrix – The following Table 1 shows a tabular representation of predictions vs. actuals, showing TP, FP, TN, and FN.

Table 1: Confusion matrix.

		Predicted Value		
		Yes	No	
Actual Value	Yes	True Positives (TP)	False Negatives (FN)	**Recall**
	No	False Positives (FP)	True Negatives (TN)	**Specificity**
		Precision	**Negative Predictive Value**	**Accuracy**

Source: Authors, (2026).

III.3 MODELLING AND COMPUTATIONAL ENVIRONMENT

A correlation matrix is a table that illustrates the correlation coefficients among variables in a dataset. It is frequently used in machine learning to find redundant or strongly correlated variables as well as to comprehend the links between various aspects. This improves interpretability by using colour gradients and annotation to visualize the matrix and is shown in Figure 3. Compared to all the independent data, the correlation coefficient between education and target variable TenYearCHD is very low and actually negative.

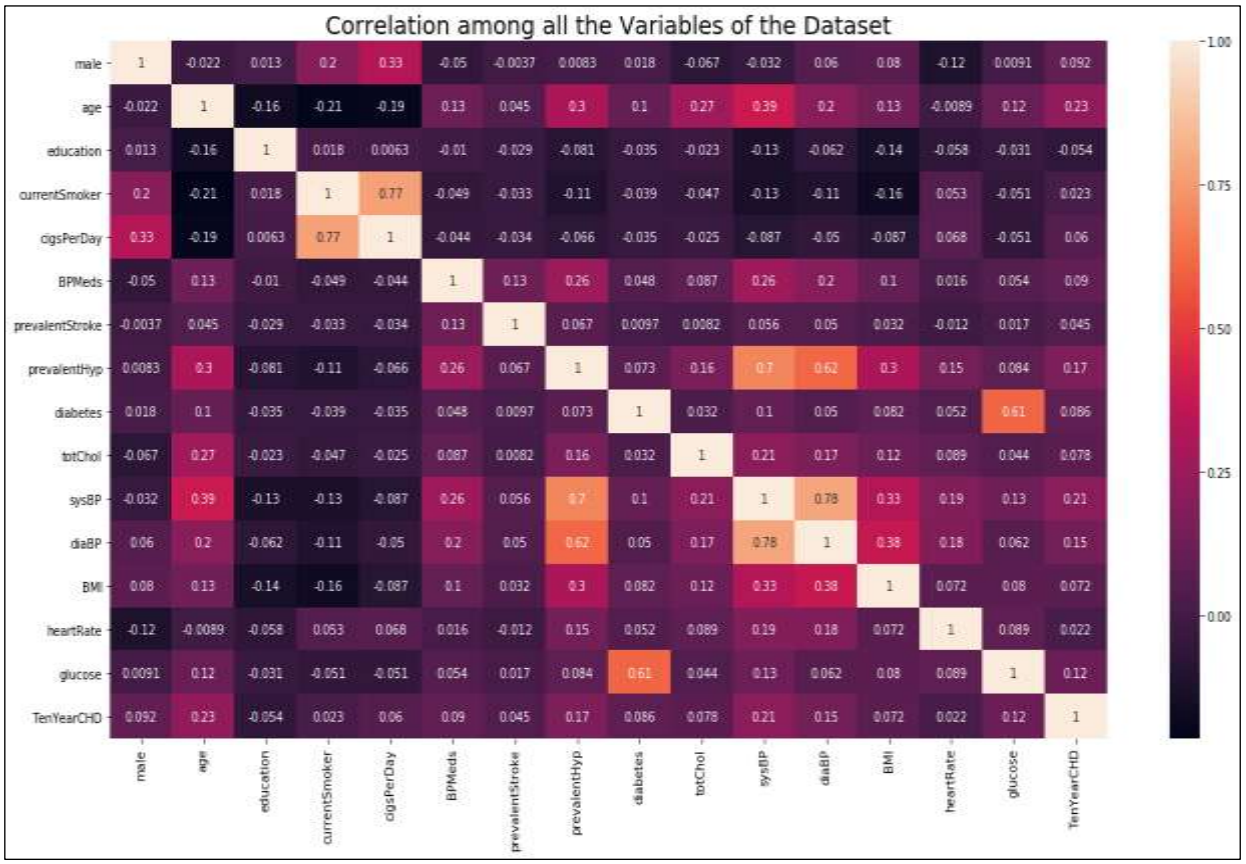


Figure 3: Correlation heatmap for dataset.
Source: Authors, (2026).

III.4 DATA ANALYSIS

III.4.1 Univariate Analysis

The study and analysis of a single variable within a dataset is known as univariate analysis. Summarizing and comprehending the properties of individual variables, such as their distribution, central tendency, dispersion, and possible anomalies, is a basic stage in exploratory data analysis (EDA). Observations from the categorical features

- BPmeds, prevalentStroke and diabetes are highly imbalanced.
- There are four levels of education whereas the rest categorical features are all binary.
- CurrentSmoker has nearly an equal number of Smokers and non-Smokers.

For the plots shown in Figure 4 observations made for numerical features are

- cigsPerDay does not have even distribution but the majority of the data is at 0
- The majority portions of the following columns lie in the range:
  - totChol: 150 to 300
  - sysBP: 100 to 150
  - diaBP: 60 to 100
  - BMI: 20 to 30
  - heartRate: 50 to 100
  - glucose: 50 to 150

The class distribution of data in the dependent feature TenYearCHD is highly imbalanced. As it were, the number of negative cases are more than the number of positive cases. This would cause class imbalance issue during the process of fitting our models. Consequently, this issue needs to be managed and handled. So, the dataset is resampled by oversampling positive cases and the classes are now balanced for model fitting. The Fig. 5 shows the distribution before and after resampling.

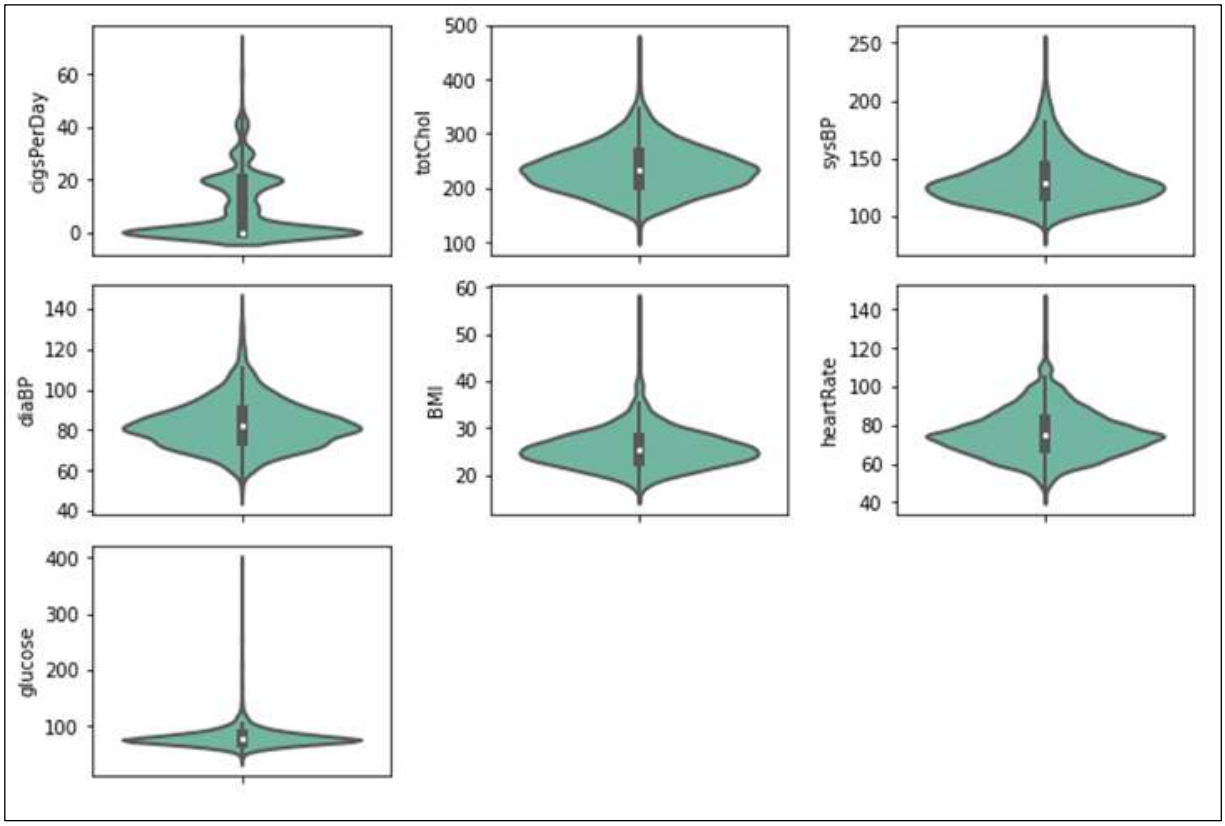


Figure 4: Data distribution in numerical features.
Source: Authors, (2026).

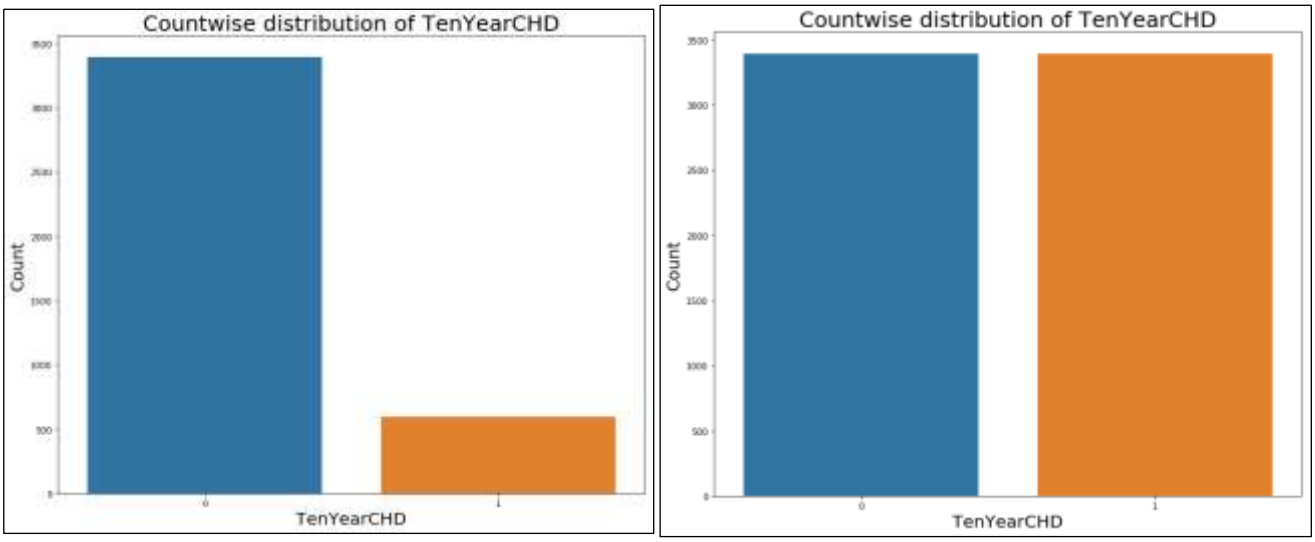


Figure 5: Before and after resampling TenYearCHD.
Source: Authors, (2026).

III.4.2 Bivariate Analysis

Bivariate analysis looks at how two variables in a dataset relate to one another. It is a crucial phase in exploratory data analysis (EDA) and aids in identifying patterns, trends, or relationships. Different methods and visualizations are employed depending on the kind of variables (numerical, categorical). From the visualization shown in Figure 6, it can also be seen that Mid-age groups between or after age 38-46 contain more currentSmokers. There is no below 32 year old currentSmokers and the maximum age of a currentSmokers is 70. The upward shift of the boxplots in Figure 7 indicates that adults had higher levels of cholesterol (bad cholesterol in general). Other observations made in this bivariate analysis are

- Non-smokers, i.e. a cigsPerDay of 0.0, are at a very low risk of developing the disease.
- Those with sysBP between 72-130 have a reduced risk of getting the disease.
- Minor relation found between higher risks of TenYearCHD with higher diaBP.
- Individuals who have diaBP of up to 80.0 are less likely to be affected by the disease.

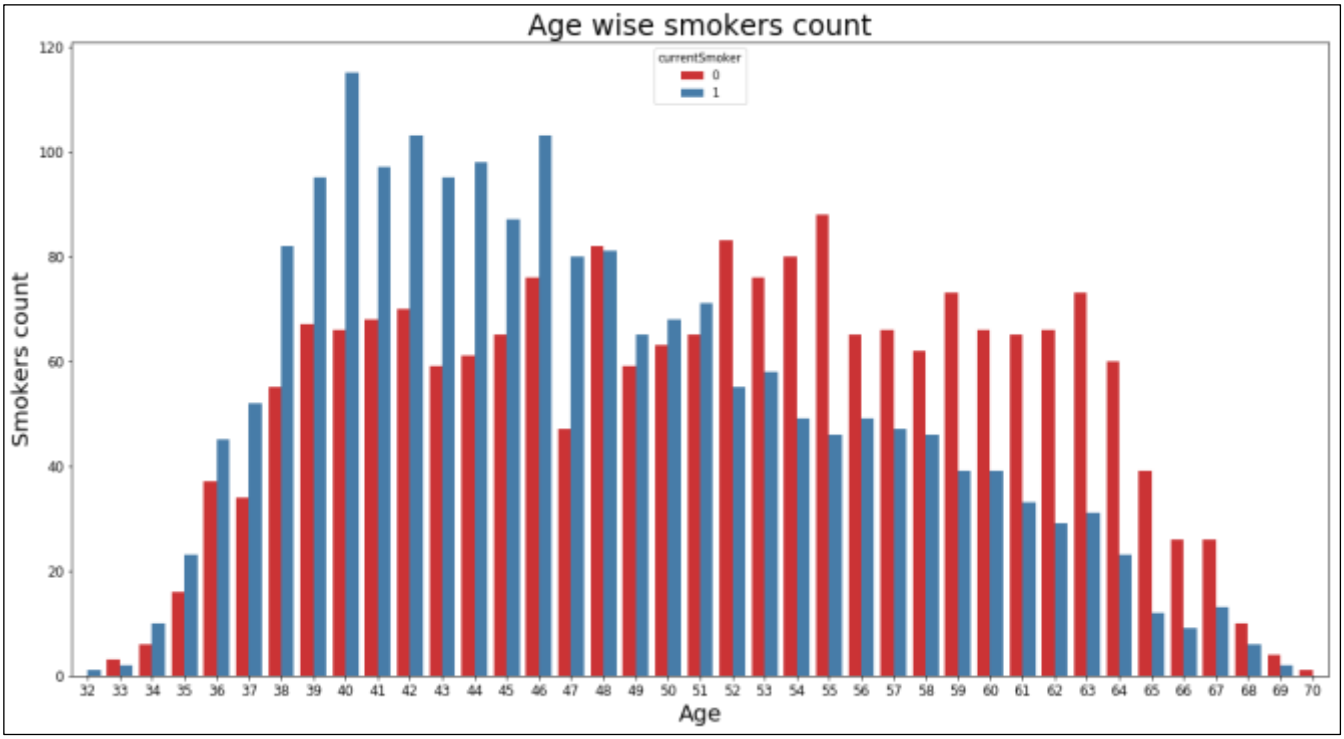


Figure 6: Age vs Smokers count.
Source: Authors, (2026).

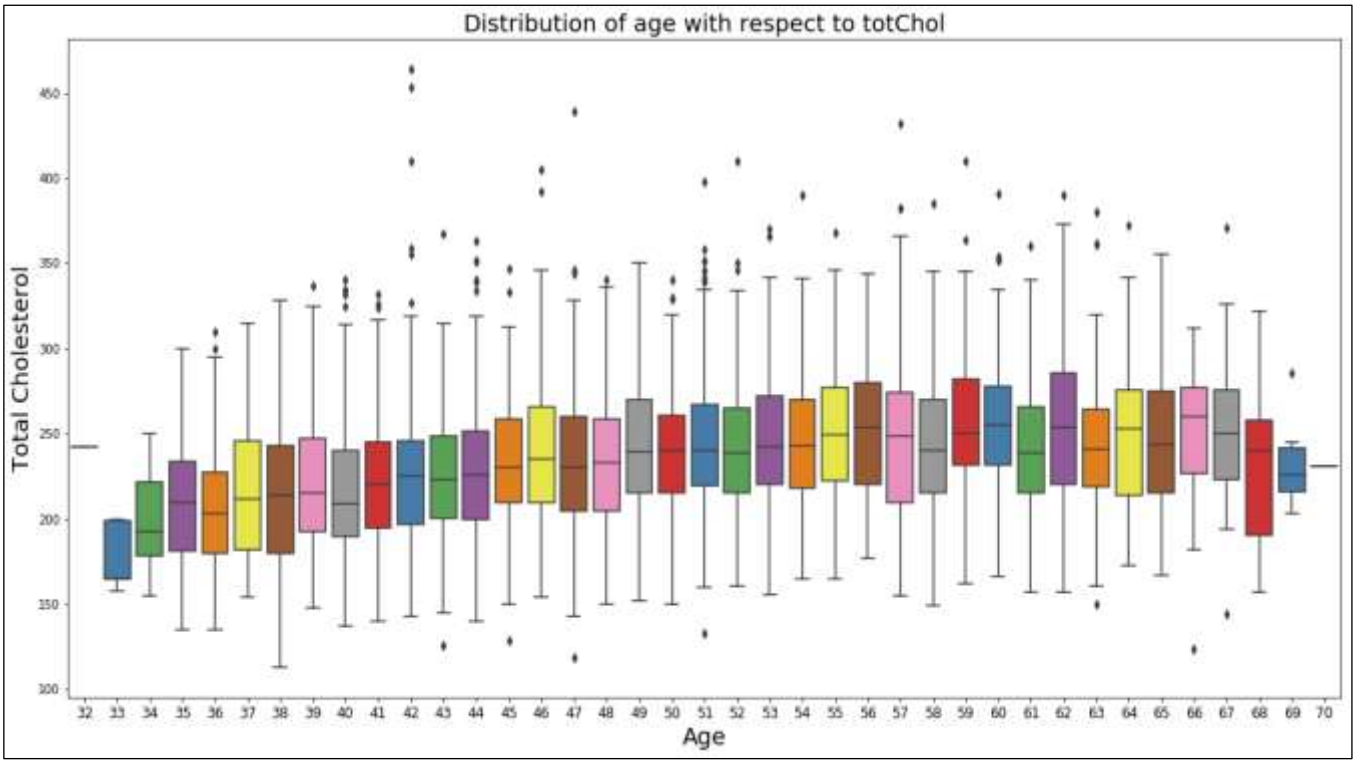


Figure 7: Distribution of Age with respect to Total Cholesterol.
Source: Authors, (2026).

III.4.3 Feature Selection

Using machine learning requires identifying and choosing the most important features contained in a dataset in order to enhance model accuracy as well as lower its complexity. It eliminates irrelevant or redundant data, which can lead to overfitting, increased computational cost, and decreased interpretability. The plot in Figure 8 shows the features and their respective chi-square test scores. The top ten features are selected for implementing the algorithms.

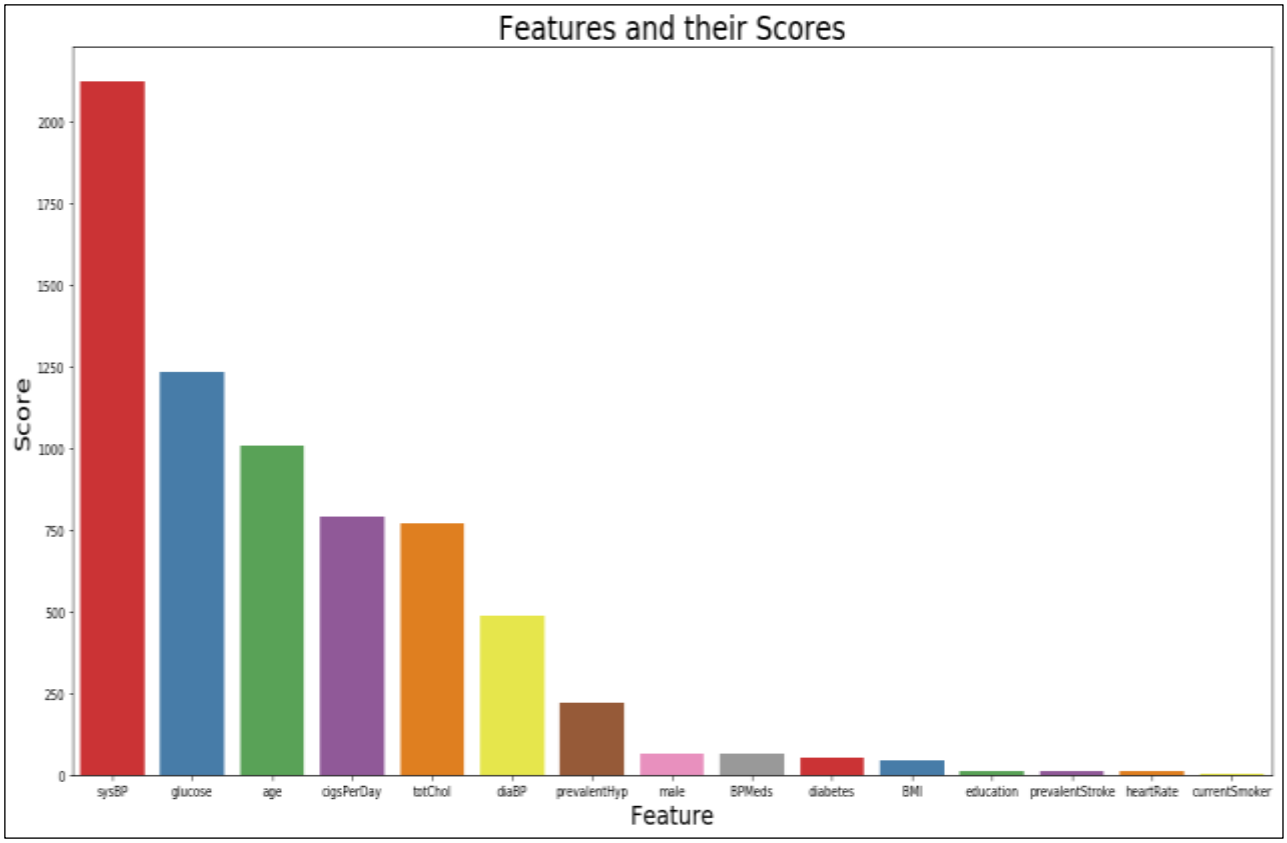


Figure 8: Features and their Chi-scores.
Source: Authors, (2026).

III.5 DATA PREPARATION

In general, the data in numerical features are of different scales and it needs to ensure that all features are on the same scale, making the model training process more stable and effective. In data science, normalization is the process of standardizing numerical data to fall into a predetermined range, usually between 0 and 1. It ensures that every feature makes an equal contribution to the analysis and keeps models from favouring features with larger scales. Min Max Scalar is applied to scale the data. The dataset is split into train and test in the ratio of 80:20.

IV. RESULTS ANALYSIS AND DISCUSSION

IV.1 CLASSIFICATION ALGORITHMS RESULTS

In this study, the dataset is trained using ten classification algorithms. Performance of these algorithms is related with respect to the metrics – Accuracy, Specificity, Precision, Recall, F1-Score and F2-Score. The Table 2 and Table 3 illustrates the results of classification algorithms on the train and test dataset respectively. From the Table 2 it is evident that, KNN Classifier, Decision Tree Classifier and XG Boost are giving 100%, 100% and 98.95% accuracy and it needs to be verified for test data also. For the test data, these algorithms are giving 93.37%, 90.86% and 91.23% accuracy. These results shows that KNN Classifier not fall under overfit problem. The confusion matrices for these three classifiers are shown in Figure 9 and KNN Classifier has made 90 wrong predictions out of 1358 samples.

Table 2: Train set results.

S.No.	Model	Accuracy	Precision	Recall	F1 Score	F2 Score
1	SGD Classifier	0.671823	0.674597	0.671823	0.670644	0.670743
2	Logistic Regression	0.677164	0.677183	0.677164	0.677163	0.677161
3	KNN Classification	1.000000	1.000000	1.000000	1.000000	1.000000
4	Naïve Bayes	0.609761	0.647355	0.609761	0.582372	0.589847
5	Decision Tree	1.000000	1.000000	1.000000	1.000000	1.000000
6	Random Forest	0.984530	0.984569	0.984530	0.984530	0.984526
7	Gradient Boosting	0.756538	0.756821	0.756538	0.756489	0.756473
8	Ada Boosting	0.689134	0.689150	0.689134	0.689134	0.689133
9	XGBoost	0.989503	0.989504	0.989503	0.989503	0.989503

Source: Authors, (2026).

Table 3: Test set results.

S.No.	Model	Accuracy	Precision	Recall	F1 Score	F2 Score
1	SGD	0.653903	0.656044	0.653903	0.652067	0.652482
2	Logistic Regression	0.667158	0.667190	0.667158	0.666999	0.667052
3	KNN Classification	0.933726	0.939146	0.933726	0.933468	0.932920
4	Naïve Bayes	0.609720	0.646893	0.609720	0.586412	0.592023
5	Decision Tree	0.908689	0.919147	0.908689	0.908010	0.907037
6	Random Forest	0.921944	0.924964	0.921944	0.921757	0.921465
7	Gradient Boosting	0.712813	0.712862	0.712813	0.712712	0.712743
8	Ada Boosting	0.679676	0.679654	0.679676	0.679656	0.679666
9	XGBoost	0.912371	0.916567	0.912371	0.912085	0.911691

Source: Authors, (2026).

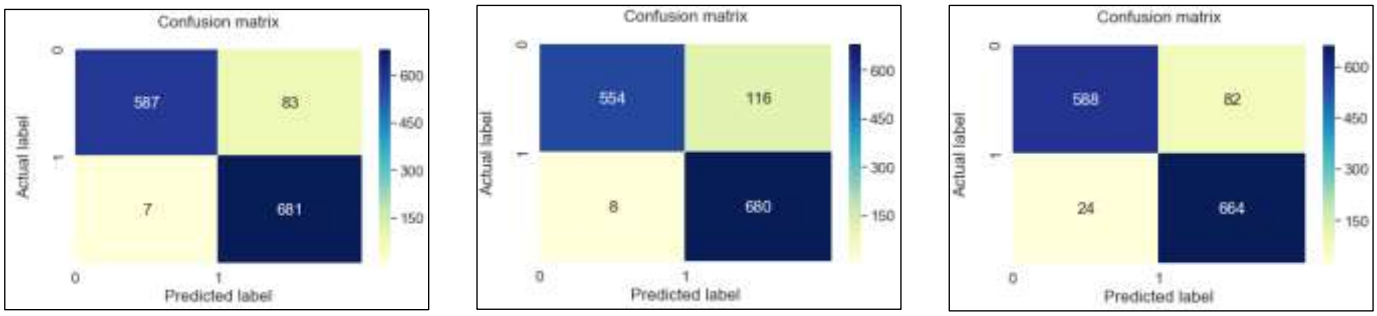


Figure 9: Confusion matrices generated for test data by a) KNN b) Decision Tree c) Random Forest.
Source: Authors, (2026).

IV.2 ENSEMBLE MODEL RESULTS

Ensemble models are a dominant tool in machine learning that combine multiple individual models to increase overall performance and robustness. By leveraging the strengths of diverse algorithms, ensemble methods can attain higher accuracy and superior generalization compared to single models. Techniques such as bagging, boosting, and stacking help reduce variance and bias, leading to more reliable predictions. The model creation process of bagging produces several models with different training data segments while boosting applies successive model optimization to tackle past prediction mistakes. A stacking method consists of creating an additional model which learns to combine predictions made by various base models. Ensemble models which use the Voting Classifier produce improved performance through the combination of multiple machine learning models. The model assembles multiple model predictions into one prediction through a majority vote for classifications and average value computation for regression problems. The two voting forms include hard voting which depends on the models' majority class predictions as well as soft voting that utilizes the individual models' averaged prediction probabilities.

Other than their practicality Ensemble models become the preferred framework for complex prediction problems which span across different application domains. Several techniques under ensemble models work together to improve predictive performance while maintaining unique methods for uniting multiple models. The current research applied three ensemble models which included bagging and stacking in addition to voting. Tables 4 and 5 exhibit the results for train and test data after the top-performing algorithms from Table 3 are chosen. XG Boost, KNN, and Random Forest classifiers are used in the suggested voting classifier, which produced an accuracy of 94.25% for test data. XG Boost, Random Forest, and Decision Tree are used as base classifiers for stacking classifier, while KNN is used as the final classifier. For both train and test data, this classifier produced 100% and 92.19% accuracy, respectively. Fig. 10 displays the confusion matrices for the three proposed ensemble models. The voting classifier outperforms the KNN classifier in terms of accuracy. Table 6 presents the results of other performance metrics and voting classifier resulted with 98.87 for ROC AUC mean.

Table 4: Train data set results for Ensemble models.

S.No.	Model	Accuracy	Precision	Recall	F1 Score	F2 Score
1	KNN Bagging	0.989871	0.989995	0.989871	0.989871	0.989856
2	Voting Classifier	0.999632	0.999632	0.999632	0.999632	0.999632
3	Stacking Classifier	1.000000	1.000000	1.000000	1.000000	1.000000

Source: Authors, (2026).

Table 5: Test data set results for Ensemble models.

S.No.	Model	Accuracy	Precision	Recall	F1 Score	F2 Score
1	KNN Bagging	0.919735	0.924276	0.919735	0.919459	0.919021
2	Voting Classifier	0.942563	0.944736	0.942563	0.942464	0.942240
3	Stacking Classifier	0.921944	0.922341	0.921944	0.921942	0.921895

Source: Authors, (2026).

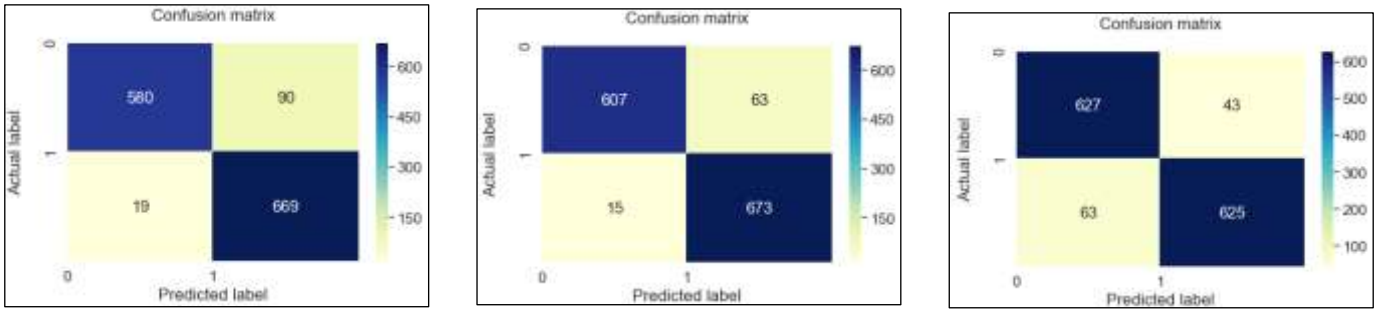


Figure 10: Confusion matrices generated for test data by a) KNN Bagging b) Voting Classifier c) Stacking Classifier.
Source: Authors, (2026).

Table 6: Accuracy metrics for Ensemble models.

S.No.	Model	ROC AUC Mean	ROC AUC STD	Accuracy Mean	Accuracy STD
1	KNN Bagging	97.23	0.99	94.88	1.38
2	Voting Classifier	98.87	0.51	96.02	1.31
3	Stacking Classifier	94.98	1.15	94.97	1.14

Source: Authors, (2026).

V. CONCLUSION

A unique ensemble approach for categorizing patients and predicting their risk for CHD was proposed in this paper. Make sure to perform pre-processing in order to comprehend the data, identify correlations between features, eliminate noisy data, balance data in two classes in the dependent feature, and extract contributing features. To manage the entire experiment with different evaluation factors such as specificity, recall, sensitivity, precision, accuracy, F-score, and confusion matrix, a Python implementation platform was used. The dataset was integrated with 4240 patients who may have a risk of coronary heart disease. KNN Classification has the best accuracy among the eight classification algorithms: SGD Classifier, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, Gradient Boosting, Ada Boosting, and XGBoost. Its training and testing accuracies are 1.0 and 0.93, respectively. The dataset is subjected to ensemble techniques like bagging, voting, and stacking classifiers. Using the best-performing classifiers – KNN, Random Forest, and Decision Tree—a voting classifier is suggested. For both train and test data, the suggested approach achieved model accuracy of 0.99 and 0.94. The ROC AUC Mean and Accuracy Mean for the work are 98.87 and 96.02, respectively.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Ch V Raghavendran, Kathi Chandra Mouli, S. Bhargavi Latha.
Methodology: Ch V Raghavendran, Kathi Chandra Mouli.
Investigation: S. Bhargavi Latha.
Discussion of results: Ch V Raghavendran, Kathi Chandra Mouli, S. Bhargavi Latha.
Writing: Review and Editing: Ch V Raghavendran, Kathi Chandra Mouli, S. Bhargavi Latha.
Resources: Kathi Chandra Mouli, S. Bhargavi Latha.
Supervision: Ch V Raghavendran, Kathi Chandra Mouli, S. Bhargavi Latha.
Approval of the final text: Ch V Raghavendran, Kathi Chandra Mouli, S. Bhargavi Latha.

VII. REFERENCES

[1] Foxwell R, Morley C, Frizelle D. Illness perceptions, mood and quality of life: a systematic review of coronary heart disease patients. J Psychosom Res. 2013 Sep;75(3):211-22. doi: 10.1016/j.jpsychores.2013.05.003.

[2] Price KJ, Gordon BA, Bird SR, Benson AC. A review of guidelines for cardiac rehabilitation exercise programmes: Is there an international consensus? Eur J Prev Cardiol. 2016 Nov;23(16):1715-1733. doi: 10.1177/2047487316657669.

[3] Candelaria D, Randall S, Ladak L, Gallagher R. Health-related quality of life and exercise-based cardiac rehabilitation in contemporary acute coronary syndrome patients: a systematic review and meta-analysis. Qual Life Res. 2020 Mar;29(3):579-592. doi: 10.1007/s11136-019-02338-y.

[4] Byrne M, Walsh J, Murphy AW. Secondary prevention of coronary heart disease: patient beliefs and health-related behaviour. J Psychosom Res. 2005 May;58(5):403-15. doi: 10.1016/j.jpsychores.2004.11.010.

[5] Bane C, Hughes CM, McElnay JC. The impact of depressive symptoms and psychosocial factors on medication adherence in cardiovascular disease. Patient Educ Couns. 2006 Feb;60(2):187-93. doi: 10.1016/j.pec.2005.01.003.

[6] Brydon L, Magid K, Steptoe A. Platelets, coronary heart disease, and stress. Brain Behav Immun. 2006 Mar;20(2):113-9. doi: 10.1016/j.bbi.2005.08.002.

[7] Juergens MC, Seekatz B, Moosdorf RG, Petrie KJ, Rief W. Illness beliefs before cardiac surgery predict disability, quality of life, and depression 3 months later. J Psychosom Res. 2010 Jun;68(6):553-60. doi: 10.1016/j.jpsychores.2009.10.004.

[8] Sebregts, E. H. W. J., Falger, P. R. J., & Bär, F. W. H. M. (2000). Risk factor modification through nonpharmacological interventions in patients with coronary heart disease. Journal of Psychosomatic Research, 48(4-5), 425–441. https://doi.org/10.1016/S0022-3999(99)00113-0.

[9] Maharani A, Sujarwoto, Praveen D, Oceandy D, Tampubolon G, Patel A (2019) Cardiovascular disease risk factor prevalence and estimated 10-year cardiovascular risk scores in Indonesia: The SMARThealth Extend study. PLoS ONE 14(4): e0215219. https://doi.org/10.1371/journal.pone.0215219.

[10] Jawalkar, A.P., Swetcha, P., Manasvi, N. et al. Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting. J. Eng. Appl. Sci. 70, 122 (2023). https://doi.org/10.1186/s44147-023-00280-y.

[11] H. Yang, Z. Chen, H. Yang and M. Tian, "Predicting Coronary Heart Disease Using an Improved LightGBM Model: Performance Analysis and Comparison," in IEEE Access, vol. 11, pp. 23366-23380, 2023, doi: 10.1109/ACCESS.2023.3253885.

[12] C. Bemando, E. Miranda and M. Aryuni, "Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms," 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), Pekan, Malaysia, 2021, pp. 232-237, doi: 10.1109/ICSECS52883.2021.00049.

[13] Dulani, H. et al. (2024). Heart Disease Prediction Using Ensemble Techniques and Explainable AI Validation. In: Dehuri, S., Cho, SB., Padhy, V.P., Shanmugam, P., Ghosh, A. (eds) Machine Intelligence, Tools, and Applications. ICMITA 2024. Learning and Analytics in Intelligent Systems, vol 40. Springer, Cham. https://doi.org/10.1007/978-3-031-65392-6_24.

[14] Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 (2020). https://doi.org/10.1007/s42979-020-00365-y.

[15] S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan and T. Zhu, "Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework," 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), Beijing, China, 2017, pp. 228-232, doi: 10.1109/ICBDA.2017.8078813.

[16] A. N. Repaka, S. D. Ravikanti and R. G. Franklin, "Design And Implementing Heart Disease Prediction Using Naives Bayesian," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 292-297, doi: 10.1109/ICOEI.2019.8862604.

[17] Du Z, Yang Y, Zheng J, Li Q, Lin D, Li Y, Fan J, Cheng W, Chen XH, Cai Y, Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation, JMIR Med Inform 2020;8(7):e17257, doi: 10.2196/17257.

[18] Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms, 16(2), 88. https://doi.org/10.3390/a16020088.

[19] Diman Hassan, Haval I. Hussein, Masoud M. Hassan, Heart disease prediction based on pre-trained deep neural networks combined with principal component analysis, Biomedical Signal Processing and Control, Volume 79, Part 1, 2023, 104019, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2022.104019.

[20] Mert Ozcan, Serhat Peker, A classification and regression tree algorithm for heart disease modeling and prediction, Healthcare Analytics, Volume 3, 2023, 100130, ISSN 2772-4425, https://doi.org/10.1016/j.health.2022.100130.

[21] Aniruddha Dutta, Tamal Batabyal, Meheli Basu, Scott T. Acton, An efficient convolutional neural network for coronary heart disease prediction, Expert Systems with Applications, Volume 159, 2020, 113408, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2020.113408.

[22] C. Beulah Christalin Latha, S. Carolin Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Informatics in Medicine Unlocked, Volume 16, 2019, 100203, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2019.100203.

[23] Mouli, K. C., Raghavendran, Ch. V., Bharadwaj, V. Y., Vybhavi, G. Y., Sravani, C., Vafaeva, K. M., … Hussein, L. (2024). An analysis on classification models for customer churn prediction. Cogent Engineering, 11(1). https://doi.org/10.1080/23311916.2024.2378877.

[24] Mouli, K. C., Raghavendran, Ch. V., Mallikarjuna Rao, Ch., Ushasree, D., Indupriya, B., Vatin, N. I., & Negi, A. S. (2024). Performance analysis of linear and non-linear machine learning models for forecasting compressive strength of concrete. Cogent Engineering, 11(1). https://doi.org/10.1080/23311916.2024.2368101

[25] B. Narsimha, Ch V Raghavendran, Pannangi Rajyalakshmi, G Kasi Reddy, M. Bhargavi and P. Naresh (2022), Cyber Defense in the Age of Artificial Intelligence and Machine Learning for Financial Fraud Detection Application. IJEER 10(2), 87-92. DOI: 10.37391/IJEER.100206.

[26] https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset/data.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 115-123. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2771

RESEARCH ARTICLE OPEN ACCESS

CARDIOVASCULAR RISK PREDICTION USING AI

Mehek Mohd Meraj Qureshi*[1], Hariram Chavan[2]

[1]B.E. degree in Information Technology from M.H. Saboo Siddik College of Engineering at Mumbai University, Mumbai. I am currently pursuing my postgraduate degree in Artificial Intelligence (AI) from K.J Somaiya Institute of Technology, Mumbai..
[2]A Professor in the Department of Information Technology at K.J Somaiya Institute of Technology, Mumbai. He holds a B.E (Computer Engineering) and M.E (Computer Science and Engineering) from WCE Sangli, Shivaji University. He holds a PhD in Computer Engineering from VJTI, Mumbai University.

[1]https://orcid.org/0009-0001-6505-5403, [2]https://orcid.org/0000-0003-1876-6971

Email: *mehek.q@somaiya.edu, hariram@somaiya.edu

ARTICLE INFO

Article History
Received: October 2, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Machine Learning
Explainable AI
Predictive Systems
Healthcare
Cardiovascular Disease.

ABSTRACT

A silent epidemic, cardiovascular diseases (CVDs) continue to be the primary reason for loss of life around the globe. This emphasizes the critical need for early-stage risk prediction models for CVDs. This implementation aims to use an existing model for CVD risk prediction, which will be AI-driven and reliable. The main objective is to use machine learning techniques on medical datasets, it also focuses on transparency of the model along with user experience to ensure its implementation in the healthcare sector. The model utilized key clinical features that were selected using distinct feature selection methods, namely analysis of variance, mutual information, and chi-square. The feature groups selected were CRF-1, CRF-2, CRF-3, CRF-4, and CRF-5, respectively. The dataset was preprocessed with SMOTE (Synthetic Minority Oversampling Technique), which helps to resolve class imbalance. All machine learning algorithms were evaluated using Stratified K-Fold Cross Validation and Randomized Search Cross Validation. The final model's interpretability was enhanced by integrating Explainable AI (XAI) techniques, especially Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). Among the algorithms tested, the Multi-Layer Perceptron (MLP) Classifier with CRF-1 achieved an AUC of 0.98 with an accuracy of 96.72%. This makes the prediction more interpretable for the healthcare professionals and patients.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The implementation focuses on the cardiovascular system present in the body. The cardiovascular system has the responsibility of circulating oxygenated blood in the body. If this cardiovascular system malfunctions, it can lead to various cardiovascular diseases (CVDs). These malfunctions arise from factors like blocked arteries or impaired heart function and thus pose significant health risks across the globe. This highlights the condition's prevalence on the rise, and it is therefore crucial to implement effective strategies to manage and reduce its impact. The rapid integration of computer science into healthcare has shown significant advancements in biomedical devices. These biomedical devices are responsible for generating vast amounts of complex medical data. The traditional manual analysis of these data is not only time-consuming but also inefficient.

Machine Learning (ML) models are capable of processing these large medical datasets and extracting meaningful insights, thus helping healthcare professionals in the decision-making process. Significant advancements have been made in recent years in machine learning to develop predictive models for cardiovascular disease by leveraging both private and publicly available datasets. Some research has successfully addressed data imbalance using techniques like SMOTE. [1] Similarly, other studies have highlighted how the integration of ML is helpful in identifying the early symptoms of cardiovascular risks. The evaluation of various algorithms has demonstrated ML's capacity to enhance predictive accuracy and assist healthcare professionals. [2]

Building on these foundations, the Cardiovascular Risk Prediction (CVRP) model aims to develop an ML model that analyzes various patient attributes and generates predictions to assist doctors and physicians as a pre-report. To ensure the reliability of our model, we assess the accuracy of multiple algorithms and fine-tune them for comparative analysis.

II. THEORETICAL REFERENCE

II.1 LITERATURE SURVEY

Numerous researchers have applied diverse ML techniques to predict CVDs. These researchers leverage the utilization of well-known datasets such as the Cleveland, UCI, and Framingham Heart Disease datasets. As detailed in Table 1, an author Hosam F. El Sofany et al. demonstrated the effectiveness of combining multiple classifiers like Decision Tree, SVM, Naïve Bayes, XGBoost, and AdaBoost for enhanced prediction. Another study in 2023, Mana Saleh Al Reshan et al. proposed a distinct method by introducing a hybrid deep learning-based architecture that used preprocessing and feature learning techniques to analyze these same public datasets.

Table 1: Related work.

Year	Author	Dataset Used	Technique used
2024	Hosam F. El-Sofany et al.[1]	Cleveland HD dataset, private HD dataset (from Egyptian specialized hospitals)	SVM, NB, DT, XGBoost, KNN, RF, Voting, AdaBoost
2023	Mana Saleh Al Reshan et al.[6]	Cleveland HD dataset	DL, Hybrid architecture, feature learning, data preprocessing
2021	Liu et al.[7]	UCI Heart Disease	RF, KNN, SVM, NB
2021	Aqsa Rahim et al [8]	Framingham, Cleveland HD and Cleveland	LR, KNN, SMOTE
2020	Hussein et al. [9]	Cleveland HD	DT, KNN
2020	Akbar et al. [10]	Cleveland HD	RF, SVM, Naïve Bayes

Source: Authors, (2026).

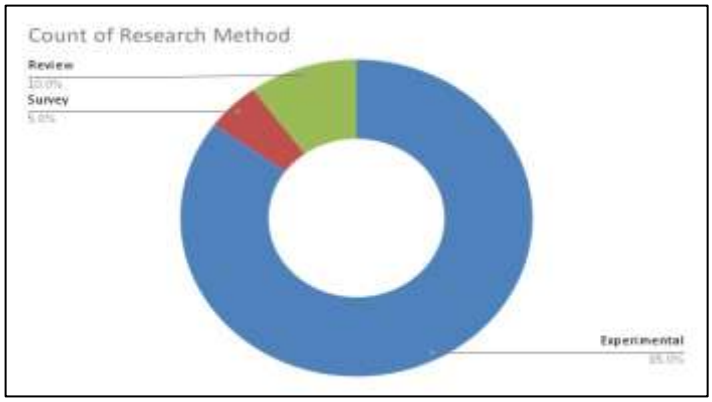


Figure 1: Distribution of different Research Methods in Research Papers
Source: Authors, (2026).

In the Figure 1, the distribution of different research methods used by researchers in 20 papers is shown. A total of 85% of the research is experimental, as shown by the largest slice. For review-based studies, it is a smaller portion of 10%. The smallest slice represents that a survey-based approach was used, accounting for 5%. Figure 2 provides a summary of ML algorithms that are most frequently used in the research papers. The size of each bar corresponds to the total number of research papers that have used a particular algorithm. This helps to identify the most prevalent algorithms in this specific domain. It also indicates which techniques are currently favored by the research community.

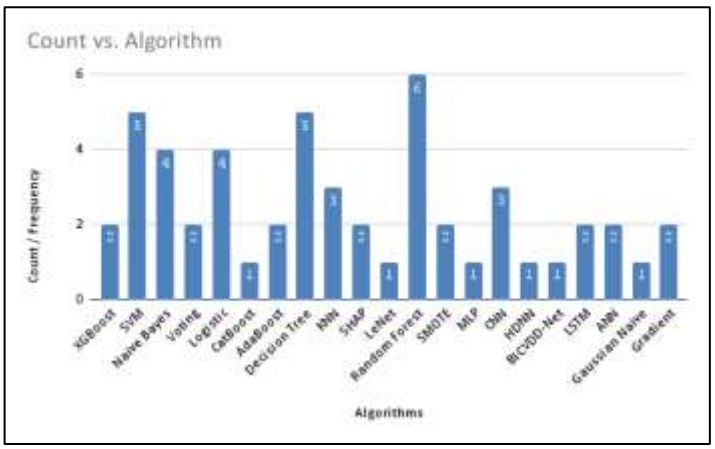


Figure 2: Frequency of Algorithms Used in Research Papers.
Source: Authors, (2026).

II.1.1 Working Principles of Classifiers

i) XGBoost / LightGBM: XGBoost/LightGBM train trees sequentially, correcting previous errors (boosting). LightGBM specifically optimizes by leaf-wise tree growth and uses Gradient-based One-Side Sampling (GOSS) for efficiency.
General form of prediction (As shown in equation 1)

$$\hat{y}_i = \sum_{k=1}^{K} f_k\,(xi), \qquad f_k \in F \tag{1}$$

where f_k are decision trees and K is the number of trees.
Objective function (As shown in equation 2):

$$\text{Obj} = \sum_{i=1}^{n} l\,(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega\,(f_k) \tag{2}$$

where l is the loss and Ω is the regularization term, which helps to avoid overfitting. ii) MLP (Multi-Layer Perceptron): It is characterized by feedforward structure and multiple layers.

Forward pass equation (Refer to equation 3):

$$h^{(l)} = \sigma\,(W^{(l)} h^{(l-1)} + b^{(l)} \tag{3}$$

where $h^{(l)}$ is the output of layer l, W is weight matrix, b is bias, and σ activation (ReLU, sigmoid, etc.).
Final prediction (As shown in equation 4):

$$\hat{y} = \sigma(W^{(L)} h^{(L-1)} + b^{(L)} \tag{4}$$

Training uses backpropagation with gradient descent (represented by equation 5):

$$W \leftarrow W - \eta \frac{\partial L}{\partial W} \tag{5}$$

where L is the loss function and η is the learning rate.

iii) Random Forest: It builds multiple independent decision trees on different bootstrapped samples of the dataset. It trains all trees independently in parallel (bagging). It is represented in equation 6.

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \tag{6}$$

where T = number of trees, $h_t(x)$ is the prediction from tree t and $\hat{y}$ is final aggregated prediction. iv) AdaBoost: It combines weak learners into a strong classifier. After each round, it increases the weights of misclassified samples, and then the next learner is allowed to focus more on those "hard" cases. The final prediction is a result of the weighted vote of all the weak learners.
Mathematical Representation:

a. Assign initial equal weights:

$$w_i = \frac{1}{N},\ i = 1,2,\ldots,N \tag{7}$$

b. Train weak learner $h_m(x)$ Compute weighted error.

$$\epsilon_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq h_m(x_i)}{\sum_{i=1}^{N} w_i} \tag{8}$$

c. Compute learner weight (importance):

$$\alpha_m = \frac{1}{2} \ln\left(\frac{1-\epsilon_m}{\epsilon_m}\right) \tag{9}$$

d. Update sample weights:

$$w_i \leftarrow w_i \cdot e^{\alpha_m I(y_i \neq h_m(x_i))} \tag{10}$$

e. Final strong classifier:

$$H(x) = sign\left(\sum_{m=1}^{M} \alpha_m h_m(x)\right) \tag{11}$$

II.1.2 CVRP System Architecture

The CVRP model is designed to analyze the risk of having CVD using various ML classification algorithms. The system architecture comprises three main layers: Frontend, Backend, and Database (As shown in Figure 3). The database used is available on the Kaggle platform. It comprises of 13 key clinical features that help to analyze the potential risks of having CVD. These attributes include a patient's age, gender, chest pain types, maximum heart rate reached, cholesterol, and more. The frontend is designed to provide a user-friendly interface that can be used to input health data by healthcare professionals and patients. This was made possible by using Gradio, it allows user to enter health data and receive an instant output about their cardiovascular risk level. The backend is responsible for the core logic. It includes data preprocessing using SMOTE, the evaluation of the ML algorithms, and the selection of the MLP Classifier model based on performance. The backend also integrates XAI techniques, which helps maintain the transparency of the model. A downloadable PDF report is generated for a summary of results.

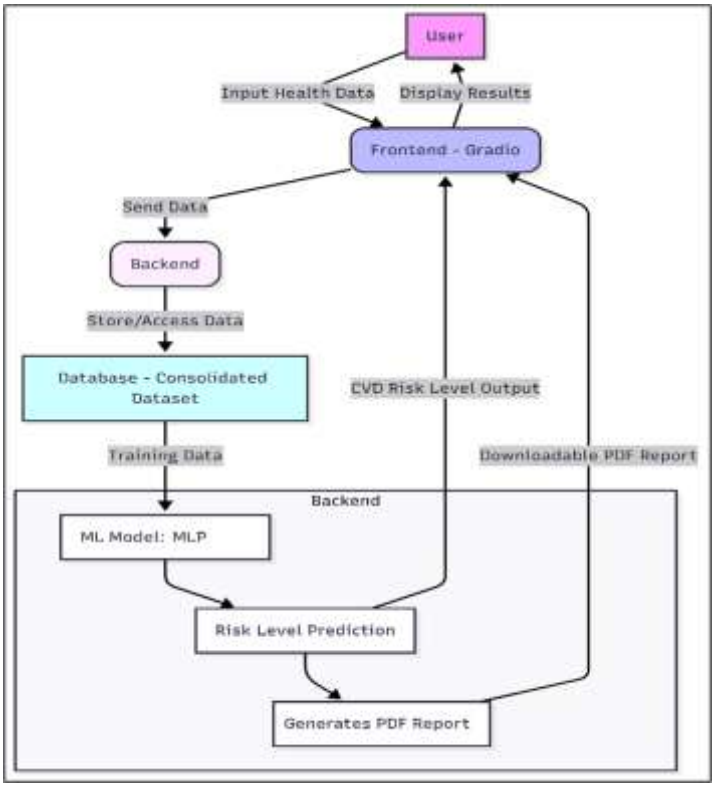


Figure 3: The CVRP System Architecture.
Source: Authors, (2026).

III. MATERIALS AND METHODS

The approach used for the CVRP model is explained in this section. The first step starts with gathering and preprocessing the dataset to prepare it for model training.

III.1 DATA PREPROCESSING

The dataset used is publicly available, it consists of 13 features, an entire summary of these features and their descriptions is mentioned in Table 2. For building any model, the most crucial task is data preprocessing. As we are dealing with medical data, data preprocessing was done with the help of tools like NumPy and pandas to ensure that we have a clean and ready dataset for training. All the missing values were handled carefully, and duplicated records were eliminated. This helps to prevent any bias or errors in the training process. The processed data was then split into two portions: 80% for training, 20% for testing.

Table 2: Features in the Dataset.

Sr. No	Attribute	Variable Name	Attribute Description	Values of Attributes
1	Age	age	Age of the person	Age in number
2	Gender	gen	The gender of the patient (0 represents female and 1 represents male)	F = 0, M = 1
3	Chest_Pain_Types	cpt	The type and severity of chest pain felt by the patient (ASY, ATA, NAP, TA)	0,1,2,3
4	Resting_BP	rbp	Denotes the patient's resting blood pressure (mm Hg)	Values between 94 and 200
5	Cholesterol	chol	Reflects the chol level measured (mg/dL)	Values between 126 and 564
6	Oldpeak	op	During physical exertion, the maximum heart rate	Values between 0 to 9.0
7	Resting_ECG	recg	Shows the patient's resting electrocardiogram. (LVH, Normal, ST)	0,1,2
8	Fasting_BS	fbs	Range of fasting blood sugar (greater than 120 mg/dL (1 = true, 0 = false))	0,1
9	Exercise_Angina	ang	Exercise-induced chest pain. (1=yes, 0=no)	0,1
10	Slope	st	ST-segment slope during peak exercise (Upsloping=2, Flat=1, and Downsloping=0), which helps assess the patient's cardiac response to stress.	0,1,2
11	Max_HR	thalach	shows the maximum heartbeat of the patient	71 to 202
12	Fluoroscopy	ca	Fluoroscopy colored main vessels count	0 to 3
13	Thallium	ths	Thallium stress (negative=0, positive=1 and inconclusive=2)	0 to 2
14	Target		This is the target or output variable in the dataset, representing binary classification. A value of "0" = lower likelihood of heart disease, while "1" = a higher risk. This classification is determined based on patterns across the other 13 input attributes.	0,1

Source: Authors, (2026).

III.2 CLASS IMBALANCE IN THE DATASET

The data was successfully preprocessed, and now we had to tackle the issue of data imbalance. Medical datasets usually have this common problem where one outcome occurs far less frequently than the other. To resolve this problem, we applied SMOTE to the training data. SMOTE balanced the dataset by creating synthetic examples for the less frequent class. This ensures that the model would not become biased towards the more common outcome (majority class). This step was crucial for building a more reliable CVRP model that provides fair and accurate predictions across all classes, especially for those at higher risk.

III.3 THE CVRP MODEL IMPLEMENTATION

To find the most accurate and consistent model for our problem, we evaluated five different machine learning algorithms, including RF, MLP, XGBoost, LightGBM, and AdaBoost. After extensive testing, the MLP Classifier was chosen according to the performance metrics. (As shown in Table 4) The final CVRP model used the MLP Classifier due to its several advantages (as shown in Table 5 and Figure 6). It learns non-linear mappings via backpropagation. It is useful for capturing interactions from features that may not be explicitly modelled in tree-based methods. It also allowed us to compare traditional machine learning with deep learning approaches in this medical domain.

III. 3.1. Model Selection and Hyperparameter Tuning

Model selection is a crucial step that helps in selecting the best model architecture or algorithm from a set of models. On the other hand, hyperparameter tuning consists of optimizing the settings of the chosen model in order to improve its performance. Feature selection methods was used to identify distinct feature that has an impact on the model's output. These methods include analysis of variance, mutual information, and chi-square. As shown in Table 3, the feature groups selected were named as Cardiac Risk Factors (CRF), a total of five groups, namely CRF-1, CRF-2, CRF-3, CRF-4, and CRF-5, were used.

Table 3: Selected Feature groups (CRF).

Selected Group of Features	Attributes
CRF-1	age, gen, chol, cpt, thalach, op, ang, ca, st, ths
CRF-2	chol, op, fbs, cpt, thalach, ang, ca, st, ths
CRF-3	gen, chol, op, recg, fbs, cpt, thalach, ang, ca, st, ths
CRF-4	op, cp, thalach, ths, ang, chol, ca, st, fbs, recg, gen, rbp
CRF-5	age, gen, chol, rbp, op, recg, fbs, cpt, thalach, ang, ca, st, ths

Source: Authors, (2026).

We employed Stratified K-Fold Cross Validation in combination with Randomized Search CV. Stratified K-Fold Cross Validation was useful because it ensured that each data fold maintained the same proportion of target classes, which is essential for medical data. Randomized Search CV efficiently found the optimal hyperparameters for each algorithm, which saved the time of manual tuning.

III.4 INTEGRATING EXPLAINABLE AI (XAI) IN CVRP MODEL

Our CVRP model goes beyond a simple yes/no prediction. It extends the output to three meaningful risk levels: low, moderate, and high. This makes the results more practical in the healthcare domain. To ensure the model is transparent and trustworthy, we integrated XAI tools. We have used SHAP values to generate a plot (Fig.4) that clearly explains the impact of each feature on a prediction. We also used LIME to generate visuals (Fig.5), which help users understand exactly how a specific prediction was made.

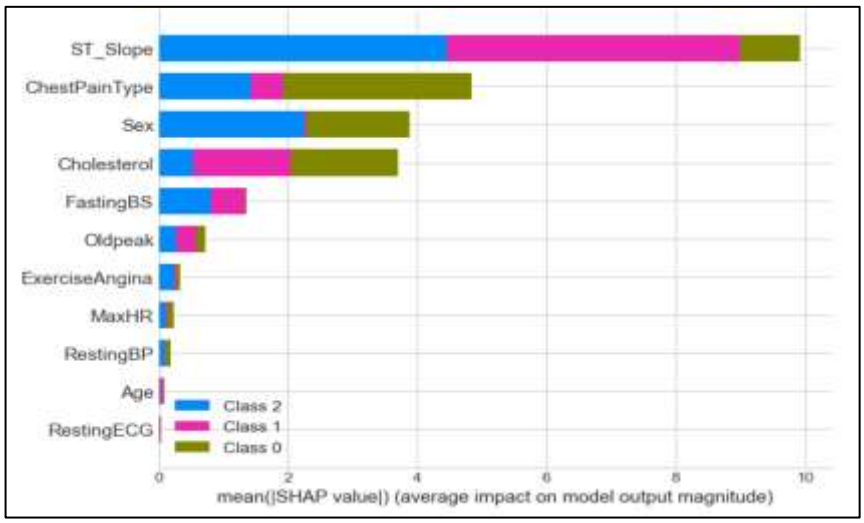


Figure 4: SHAP Bar Plot (Feature Importance).
Source: Authors, (2026).

The bar plot in Figure 5 shows the overall feature importance for the model. It helps understand that, on average, features like ST_Slope and ChestPainType are the most effective features for the predictions.

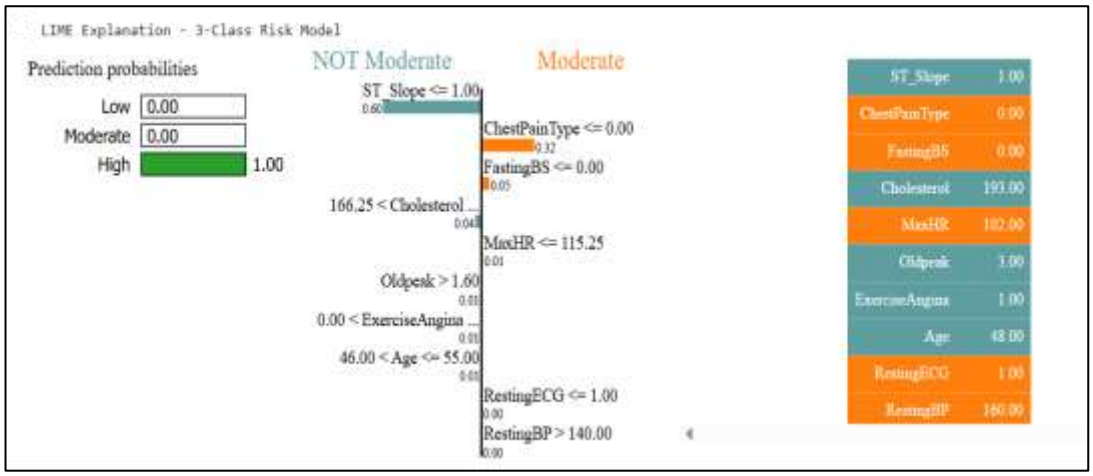


Figure 5: LIME explanation for a patient.
Source: Authors, (2026).

Figure 5: This LIME plot helps analyze which feature value positively or negatively influenced the model's result for an individual patient. In the final step, we saved the best model and built a user-friendly platform for healthcare professionals and patients. This web-based app was built using Gradio. It allows to enter health data and instantly the app will give an output about the level of cardiovascular risk. It even generates a downloadable PDF report that summarizes the results. This made the tool both intelligent and efficient to use for patients and healthcare professionals.

IV. RESULTS AND DISCUSSIONS

For the CVRP model, it is crucial to understand the model's ability to generalize across diverse patient data and detect CVDs. This can be done using performance evaluation techniques. These techniques help evaluate the accuracy and effectiveness of a model. It also helps to identify the areas for improvement (As shown in Table 4).

IV.1 PERFORMANCE EVALUATION TECHNIQUES

For evaluation of the CVRP model's performance, the following parameters were utilized:

1. **Accuracy:** It represents the percentage of correctly predicted cases of CVDs, which shows the overall effectiveness of the model. It is given as in Eq.12

$$\text{Accuracy} = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{TP+TN+FN+FP} \tag{12}$$

2. **Precision:** This represents the proportion of true positive (TP) results out of all the positive predictions. This signifies the reliability of the model's positive outcomes. It is given as in Eq.13

$$\text{Precision} = \frac{True\ Positive\ (TP)}{TP+FP} \tag{13}$$

3. **Recall (Sensitivity):** Recall is used to identify the proportion of all actual cases of CVDs that were classified correctly as positives by the model. It is given as in Eq.14

$$\text{Recall} = \frac{True\ Positive\ (TP)}{TP+FN} \tag{14}$$

4. **F1 Score:** A consolidation of precision and recall is F1 score. It is given as in Eq.15
5.

$$\text{F1} - \text{Score} = 2\frac{Precision*Recall\ (Sensitivity)}{Precision+Recall\ (Sensitivity)} \tag{15}$$

6. **Execution Time:** It denotes the time required to process the dataset and generate predictions.

7. **ROC-AUC Score:** This helps to understand the ability of the model to distinguish between the healthy patients from the at-risk patients. As shown in Figure 7.

8. **Cost:** It is the expenses associated with data storage, processing, and cloud services utilized during model implementation.

Table 4: Performance evaluation results of ML techniques using CRFs.

ML Algorithm		MLP	RF	XGBOOST	LIGHTGBM	ADABOOST
CRF-1	Accuracy %	96.72	93.44	95.89	95.08	93.44
	Sensitivity %	97.00	94.00	96.00	95.00	94.00
	F1 score %	97.00	94.00	96.00	95.00	94.00
CRF-2	Accuracy %	91.08	91.08	95.08	95.08	91.80
	Recall %	94.00	88.00	91.00	91.00	91.00
	F1 score %	92.00	92.00	95.00	95.00	92.00
CRF-3	Accuracy %	93.44	91.08	95.90	95.08	91.08
	Recall %	94.00	88.00	94.00	91.00	94.00
	F1 score %	94.00	92.00	96.00	95.00	92.00
CRF-4	Accuracy %	93.44	93.44	95.08	95.08	93.44
	Recall %	94.00	91.00	94.00	91.00	94.00
	F1 score %	94.00	94.00	95.00	95.00	94.00
CRF-5	Accuracy %	91.08	90.16	95.08	95.08	95.08
	Recall %	91.00	88.00	91.00	91.00	94.00
	F1 score %	92.00	90.00	95.00	95.00	95.00

Source: Authors, (2026).

Table 5: Performance results of MLP Classifier using CRF-1 and SMOTE.

ML Algorithm	Accuracy	Precision	Sensitivity	F1 score	AUC
MLP	96.72%	97.00%	97.00%	97.00%	0.98

Source: Authors, (2026).

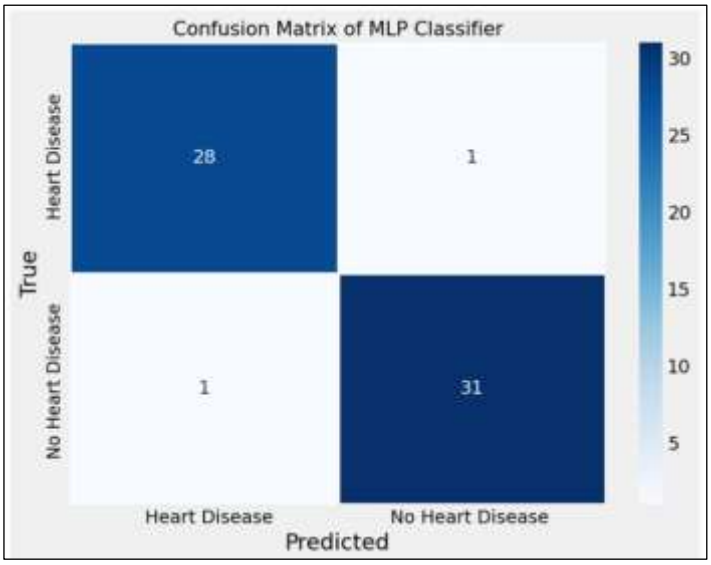


Figure 6: Confusion matrix of MLP Classifier using CRF-1 and SMOTE.
Source: Authors, (2026).

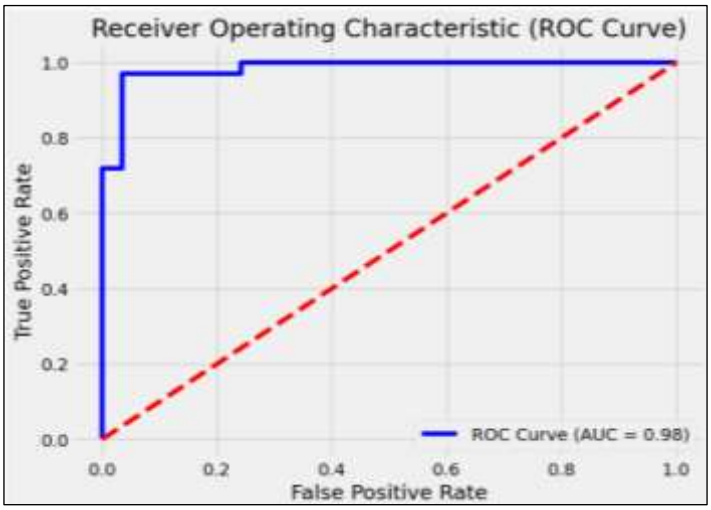


Figure 7: AUC and ROC curve of MLP Classifier using CRF-1 and SMOTE.
Source: Authors, (2026).

IV. 2. LIMITATIONS AND FUTURE SCOPE

The CVRP model's aim was to enhance the predictive accuracy of CVD prediction by integrating machine learning techniques. It will be a reliable platform that will help healthcare professionals and patients. As there were dataset limitations, the model couldn't reach a comparable accuracy with the existing solutions. Future plans include integrating datasets from different sources, including medical imaging, along with telehealth services. This will further improve the CVRP model's accuracy and effectiveness. This system can also be enhanced by analyzing and defining the stages of CVDs, which will help identify the risk of CVDs in early stages.

IV. 3. D]ISCUSSIONS

The CVRP model used MLP Classifier which achieved a high accuracy of 96.72% with AUC of 0.98. The model has a user-friendly frontend, which makes it convenient to identify CVD risk by healthcare professionals and patients. The system also maintains transparency and interpretability by using XAI techniques. It makes predictions more understandable, thus making it effective in real-time usage. This system aspires to develop a high-confidence model for early risk detection and management of CVDs.

V. CONCLUSIONS

The CVRP model's aim was to enhance the overall predictive accuracy of CVD prediction. This was achieved by using a publicly available dataset and integrating machine learning techniques. CVRP will be a reliable platform that will help healthcare professionals assess patient health reports. ML algorithms are evaluated and selected depending on dataset characteristics to improve the model reliability and effectiveness in cardiovascular risk prediction. CVRP helps prioritize user experience while addressing privacy concerns and ensuring accessibility for healthcare providers. The integration of Explainable AI (XAI) introduces an approach to improving model transparency and interpretability, making cardiovascular risk predictions more reliable and understandable.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Mehek Mohd Meraj Qureshi, Hariram Chavan.
Methodology: Mehek Mohd Meraj Qureshi, Hariram Chavan.
Investigation: Mehek Mohd Meraj Qureshi, Hariram Chavan.
Discussion of results: Mehek Mohd Meraj Qureshi, Hariram Chavan.
Writing – Original Draft: Mehek Mohd Meraj Qureshi, Hariram Chavan.
Writing – Review and Editing: Mehek Mohd Meraj Qureshi, Hariram Chavan.
Resources: Mehek Mohd Meraj Qureshi, Hariram Chavan.
Supervision: Mehek Mohd Meraj Qureshi, Hariram Chavan.
Approval of the final text: Mehek Mohd Meraj Qureshi, Hariram Chavan.

VII. REFERENCES

[1] H. F. El-Sofany, "Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques," in IEEE Access, vol. 12, pp. 106146-106160, 2024, doi: 10.1109/ACCESS.2024.3437181.

[2] Z. Alom, M. A. Azim, Z. Aung, M. Khushi, J. Car, and M. A. Moni, ''Early stage detection of heart failure using machine learning techniques,'' in Proc. Int. Conf. Big Data, IoT, Mach. Learn., Cox's Bazar, Bangladesh, 2021, pp. 23–25.

[3] V. Vision Paul and J. A. I. S. Masood, "Exploring Predictive Methods for Cardiovascular Disease: A Survey of Methods and Applications," in IEEE Access, vol. 12, pp. 101497-101505, 2024, doi: 10.1109/ACCESS.2024.3430898.

[4] H. Khan, N. Javaid, T. Bashir, M. Akbar, N. Alrajeh and S. Aslam, "Heart Disease Prediction Using Novel Ensemble and Blending Based Cardiovascular Disease Detection Networks: EnsCVDD-Net and BlCVDD-Net," in IEEE Access, vol. 12, pp. 109230-109254, 2024, doi: 10.1109/ACCESS.2024.3421241.

[5] M. S. A. Reshan, S. Amin, M. A. Zeb, A. Sulaiman, H. Alshahrani and A. Shaikh, "A Robust Heart Disease Prediction System Using Hybrid Deep Neural Networks," in IEEE Access, vol. 11, pp. 121574-121591, 2023, doi: 10.1109/ACCESS.2023.3328909.

[6] Y. Liu, X. Li, and J. Ren, ''A comparative analysis of machine learning algorithms for heart disease prediction,'' Comput. Methods Programs Biomed., vol. 200, Nov. 2021, Art. no. 105965.

[7] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim and A. W. Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," in IEEE Access, vol. 9, pp. 106575-106588, 2021, doi: 10.1109/ACCESS.2021.3098688.

[8] N. S. Hussein, A. Mustapha, and Z. A. Othman, ''Comparative study of machine learning techniques for heart disease diagnosis,'' Comput. Sci. Inf. Syst., vol. 17, no. 4, pp. 773–785, 2020.

[9] S. Akbar, R. Tariq, and A. Basharat, ''Heart disease prediction using different machine learning approaches: A critical review,'' J. Ambient Intell. Humanized Comput., vol. 11, no. 5, pp. 1973–1984, 2020.

[10] A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452-457, DOI: 10.1109/ICE348803.2020.9122958.

[11] D. Krishnani, A. Kumari, A. Dewangan, A. Singh and N. S. Naik, "Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 2019, pp. 367-372, DOI: 10.1109/TENCON.2019.8929434

[12] M. S. Raja, M. Anurag, C. P. Reddy and N. R. Sirisala, "Machine Learning-Based Heart Disease Prediction System," 2021 International Conference on Computer Communication and Informatics (ICCCI), 2021, pp. 1-5, DOI:
10.1109/ICCCI50826.2021.9402653.

[13] R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 302-305, DOI:
10.1109/ICESC48915.2020.9155586.

[14] S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019, pp. 1-5, DOI: 10.1109/ICIICT1.2019.8741465.

[15] Pabitra Kumar Bhunia, Arijit Debnath, Poulami Mondal, Monalisa D E, Kankana Ganguly, Pranati Rakshit, "Heart Disease Prediction using Machine Learning," International Journal Of Engineering Research & Technology (IJERT) NCETER – 2021 (Volume 09 – Issue 11), DOI: 10.17577/IJERTCONV9IS11071.

[16] Himanshu Sharma, M. A. Rizvi, "Prediction of Heart Disease Using Machine Learning Algorithms: A Survey," 2017 International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), pp. 99-104, DOI: 0.17762/ijritcc.v5i8.1175.

[17] P. Kumar, S. K. Sahu, "Predicting Heart Disease with Machine Learning Techniques: A Review," 2023 IEEE ICCIS, pp. 100-104, DOI: 10.1109/ICCCIS.2023.950047.

[18] V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 177-181, DOI: 10.1109/ICACCCN51052.2020.9362842.

[19] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh and V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT),
2018, pp. 1-7, DOI: 10.1109/ICCTCT.2018.8550857.

[20] S. Akbar, R. Tariq, and A. Basharat, ‘‘Heart disease prediction using different machine learning approaches: A critical review,’’ J. Ambient Intell. Humanized Comput., vol. 11, no. 5, pp. 1973–1984, 2020.

[21] M. Liu, X. Sun, Y. Liu, X. Yang, Y. Xu, and X. Sun, ‘‘Deep learningbased prediction of coronary artery disease with CT angiography,’’ Jpn. J. Radiol., vol. 38, no. 4, pp. 366–374, 2020.

[22] A. Shoukat, S. Arshad, N. Ali, and G. Murtaza, ‘‘Prediction of cardiovascular diseases using machine learning: A systematic review,’’ J. Med. Syst., vol. 44, no. 8, p. 162, Aug. 2020.

[23] N. Khandadash, E. Ababneh, and M. Al-Qudah, ‘‘Predicting the risk of coronary artery disease in women using machine learning techniques,’’ J. Med. Syst., vol. 45, p. 62, Apr. 2021.

[24] H. El-Sofany, S. A. El-Seoud, O. H. Karam, Y. M. Abd El-Latif, and I. A. T. F. Taj-Eddin, ‘‘A proposed technique using machine learning for the prediction of diabetes disease through a mobile app,’’ Int. J. Intell. Syst., vol.
2024, pp. 1–13, Jan. 2024.

[25] V. Bento, M. Kohler, P. Diaz, L. Mendoza, and M. A. Pacheco, ‘‘Improving deep learning performance by using explainable artificial intelligence (XAI) approaches,’’ Discover Artif. Intell., vol. 1, no. 1, pp. 1–11, Dec. 2021.

[26] D. Y. Omkari and K. Shaik, "An Integrated Two-Layered Voting (TLV) Framework for Coronary Artery Disease Prediction Using Machine Learning Classifiers," in IEEE Access, vol. 12, pp. 56275-56290, 2024, doi: 10.1109/ACCESS.2024.3389707.

[27] J. J. Gabriel and L. Jani Anbarasi, "Accurate Cardiovascular Disease Prediction: Leveraging Opt_hpLGBM With Dual-Tier Feature Selection," in IEEE Access, vol. 12, pp. 142427-142448, 2024, doi: 10.1109/ACCESS.2024.3470537.

[28] Available at https://www.geeksforgeeks.org/machine-learning-with-python/

[29] Available at https://www.freecodecamp.org/learn/machine-learning-with-python/

[30] Available at https://www.deeplearning.ai/courses/machine-learning-in-production/

[31] Available at https://medium.com/@darshanaweerasooriya11/introduction-to-artificial-intelligence-machine-learning-deep-learning-81b6b29d98d7

[32] World Health Organization. Cardiovascular Diseases (CVDs). Accessed: May 5, 2023. [Online]. Available: https://www.afro.who.int/ health-topics/cardiovascular-diseases.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 124-132. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2800

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

DIFFERENTIABLEHMM: NEURAL DIFFERENTIABLE HIDDEN MARKOV MODEL

Arbia Boudaoud*[1], Salheddine Kabou[2]

[1]Faculty of Mathematics and Computer Science,African University Ahmed Draia,LDDI Laboratory , Adrar, Algeria
[2]Department of exact sciences,Higher Normal School of Bechar, Bechar, Algeria

[1]https://orcid.org/0009-0007-3735-4477, [2]https://orcid.org/0000-0002-1423-7215

Email: *bouda.arbia@univ-adrar.edu.dz, kabou.salheddine@ensbechar.dz

ARTICLE INFO

Article History
Received: October 9, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Hidden Markov,
LSTM,
SequenceModeling,
Expectation–Maximization,
Clinical Data.

ABSTRACT

Modeling clinical time series demands interpretable patient states alongside the ability to capture long-range dependencies. Classical HMMs provide discrete, clinically meaningful states but ignore distant history; RNNs capture rich temporal patterns but act as black boxes. We introduce DifferentiableHMM, a unified, end-to-end differentiable probabilistic model, not a simple stacking of an HMM and a neural net. An LSTM encoder produces history-aware representations that condition both transition probabilities and GMM emission mixture weights via a small MixtureSelector network; transitions, emissions and the selector are optimized jointly by backpropagating through EM-like inference, producing a single coherent model in which inference and learning reinforce one another. On a MIMIC-III ICU benchmark, DifferentiableHMM reduces sequence negative log-likelihood by 40%, increases one-step R^2 by 5%, and lowers calibration error by 24% versus HMM and LSTM baselines. Discrete latent states retain clinical interpretability (mutual information = 0.824) and ablations show the performance gains disappear if components are trained separately. DifferentiableHMM therefore reconciles interpretability, calibrated uncertainty, and neural expressivity in a principled framework for healthcare sequence modeling.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Modeling patient trajectories in intensive care units presents a dual challenge: capturing interpretable clinical states while leveraging long-range temporal dependencies. Traditional Hidden Markov Models (HMMs) provide a clear, discrete representation of latent physiological regimes, enabling clinicians to associate each state with recognizable patterns such as “stable,” “deteriorating,” or “recovery.” However, their first-order, memoryless transition structure limits the ability to incorporate extended patient history, often leading to oversimplified dynamics. On the other hand, recurrent neural networks—particularly long short-term memory (LSTM) architectures—excel at learning complex temporal patterns over long horizons but yield continuous latent representations that lack direct interpretability and clinical semantic grounding. To address these limitations, we introduce DifferentiableHMM, a hybrid framework that unites HMM interpretability with RNN expressivity. Our key contributions are: LSTM-based history encoder for aware transition matrix, Fully Differentiable Inference and Bayesian-Informed Initialization.

To demonstrate the effectiveness of DifferentiableHMM, we evaluate on the MIMIC-III benchmark of de-identified ICU stays. Because clinical time-series are highly sensitive, developing and evaluating models on datasets like MIMIC-III also requires careful attention to privacy-preserving dynamic data publishing techniques (e.g., anonymization, sanitization, and synthetic data generation) to enable safe sharing of patient trajectories [1]. Compared to classical HMMs and pure LSTM baselines. Ablation studies confirm that each component (history conditioning, differentiable inference, and informed initialization) is essential to these gains. DifferentiableHMM thus offers a principled, high-fidelity, and interpretable approach to sequence modeling with broad potential applications in healthcare analytics and beyond. The remainder of this paper is organized as follows. Section 2 surveys related work. Section 3 explore the key concepts and fundamental components of the DifferentiableHMM method. Section 4 outlines the experimental protocol and discusses the resulting performance. Finally, Section 5 concludes the paper and outlines directions for future research.

II. RELATED WORK

The fusion of neural sequence models with Hidden Markov Models (HMMs) has been pursued in diverse domains—ranging from video action segmentation and speech processing to healthcare risk prediction—but existing hybrids universally either fix the HMM transitions or learn neural and probabilistic components in separate stages. Below, we summarize key prior contributions and contrast them with our end-to-end DifferentiableHMM.

II.1 VIDEO ACTION SEGMENTATION

[2] tackle weakly supervised temporal action segmentation in long, untrimmed videos. Their hybrid RNN–HMM framework first employs a bidirectional RNN to produce frame-level 'action scores' (emission probabilities), then applies an HMM with a fixed transition matrix to enforce temporal consistency and infer segment boundaries. [3] similarly use a BLSTM RNN to generate precise phonetic emission probabilities in speech tasks, which an HMM then decodes with static transitions to impose temporal and phonotactic constraints. In both cases, the neural component enhances the emissions, but does not influence the transition dynamics.

II.2 SEMI-MARKOV AND DURATION MODELING

To model explicit state durations, [4] introduce the Recurrent Hidden Semi Markov Model (RHSMM), where an RNN predicts dwell-time distributions before EM estimates semi Markov parameters. While more expressive than standard HMMs, RHSMM still decouples neural training from transition learning. [5]. apply a hybrid NN–HMM for automatic story segmentation, using a feed-forward network for emission posteriors and a fixed HMM for temporal smoothing, again without enabling neural influence on transitions.

II.3 SPEECH SYNTHESIS AND RECOGNITION

[6] pioneered CNN–HMM hybrids for speech recognition by replacing GMM emissions with convolutional neural networks, but retained a transition matrix learned solely via EM. Zen [7] reviews the evolution from HMM to LSTM RNN acoustic models in statistical parametric speech synthesis, showing that recurrent networks markedly improve emission quality while leaving transitions intact. [8] present NeuralHMM, parameterizing both transitions and emissions with deep networks for attention-free TTS; however, their training does not backpropagate through the HMM inference, and transitions are optimized only for frame-level synthesis.

II.4 HEALTHCARE AND RISK PREDICTION

[9] propose a hybrid GRU–attention–HMM for ICU mortality and length of stay prediction. A GRU with attention produces time-point saliency weights that guide a fixed HMM transition matrix and a Mixture Density Network emission layer; training proceeds in two stages (neural likelihood then Baum–Welch), with no joint optimization. In contrast to all the above, our DifferentiableHMM fully merges an LSTM based history encoder into the transition network and employs a differentiable Expectation–Maximization algorithm that back propagates through the forward–backward recursions. This unified, end to end approach yields discrete, interpretable latent states whose dynamics are both history aware and jointly optimized with emissions—capabilities absent in prior work.

Table 1: Comparative overview of sequence modeling approaches.

Methods	Area of Concern	Neural Component	Transition Parameterization	Joint E/M-Step Differentiability	HistoryConditioning	Objective
Kuehne et al.[2]	action segmentation	BLSTM	Fixed HMM	No	✗	Discriminative
Sun et al. [3]	speech recognition	BLSTM	Fixed HMM	No	✗	Discriminative
Dai et al. [4]	Sequence modeling	RNN duration predictor	Semi-Markov fixed (EM)	No	✗	Generative (duration)
Yu et al. [5]	Story segmentation	Feed-forward NN emissions	Fixed (EM)	No	✗	Discriminative
Abdel-Hamid et al. [6]	Speech recognition	CNN emissions	Fixed (EM-learned)	No	✗	Discriminative
Zen [7]	Speech synthesis	LSTM-RNN emissions	Fixed (Baum–Welch)	No	✗	Generative (TTS)
Mehta et al. [8]	Text-to-speech (TTS)	Deep nets for both transitions & emissions	Neural-parametrized, but trained separately	No	✗	Generative (TTS)
Liu et al. [9]	Healthriskprediction	GRU + attention + MDN emissions	Fixed HMM matrix guided by attention	No (two-stage training)	✗	Discriminative

Source: Authors, (2026).

III. METHODOLOGY

III.1 BACKGROUND

III.1.1 Hidden Markov Models (HMM)

A Hidden Markov Model (HMM) [10] is a doubly stochastic process with a discrete latent state sequence $\{s_t\}$ and observations $\{x_t\}$ .At each time step:

Transition Matrix: $A_{ij} = P(s_t = j|s_{t-1} = i)$ probability of moving between states

Emission Probabilities: $B_j(o_t) = P(o_t|s_t = j)$ likelihood of an observation given a hidden state

HMMs enjoy interpretable, discrete "modes" of behavior and exact inference algorithms. However, their first-order, memoryless transitions and fixed transition matrix limit their ability to capture long-range dependencies in complex time series.

III.1.2 Expectation–Maximization for HMMs

Parameter estimation in HMMs is traditionally performed via the Baum–Welch algorithm, an instance of the Expectation–Maximization (EM) [11] framework:

E-step: Compute the posterior state marginals and pairwise transition counts given current parameters.

M-step: Update transition and emission parameters to maximize the expected complete-data log-likelihood.

While EM guarantees non-decreasing likelihood, it treats the transition matrix as a free parameter, preventing integration with continuous function approximators and end-to-end gradient flows.

III.1.3 Recurrent Neural Networks for Sequence Modeling

Recurrent neural architectures—especially LSTMs—[12] have achieved strong predictive performance on sequential data by learning rich, continuous hidden representations:

$$h_t = LSTM(x_t, h_{t-1}), \hat{x}_{t+1} = f_\theta(h_t)$$

However, purely neural models lack discrete latent states, which are often desirable for interpretability and downstream decision support in domains such as healthcare.

III.2 DIFFERENTIABLEHMM MODEL

Figure 1 illustrates the overall DifferentiableHMM architecture. The observed time series is first passed through an LSTM history encoder; its hidden representation conditions two modules — a transition network that produces history-aware transition probabilities over discrete latent states, and a MixtureSelector that outputs logits for the GMM components used by each state's emission model. Emissions are represented by per-state GMMs, and EM-like inference is unrolled and differentiated so that transitions, mixture weights, and GMM parameters are optimized jointly via backpropagation. The result is a single, end-to-end trainable model that provides history-dependent predictions while preserving interpretable discrete states.

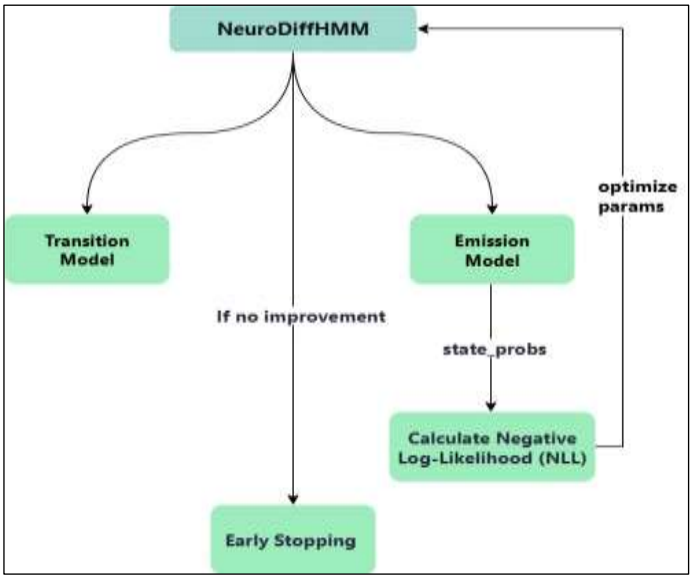


Figure 1: Overall architecture of DifferentiableHMM.
Source: Authors, (2026).

III.2.1 Step 1: Data Preprocessing

Objective: Transform raw ICU time series into structured sequences

- Excluding non-informative columns (e.g., icustay_id, charttime).

- Standardizing all features using the mean and standard deviation computed across the training set.
- Segmenting each patient's time-series into fixed-length sequences (length 50) for model input.

III.2.2 Step 2: Model Design (DifferentiableHMM)

The Differentiable Hidden Markov Model (DifferentiableHMM) is a neural extension of the classical HMM, tailored for modeling complex ICU time-series data. Unlike traditional HMMs, which rely on fixed transition and emission probabilities, the DifferentiableHMM incorporates neural network components to enable end-to-end differentiable learning and greater modeling flexibility. The model consists of a learnable transition network, typically parameterized by an LSTM-based history encoder, which captures temporal dependencies and patient-specific context. Emission distributions are modeled as mixtures of Gaussians, with mixture weights and parameters learned via backpropagation. Initialization of emission means leverages k-means clustering for stability, while mixture variances are set based on data statistics. The model supports regularization through entropy and L2 penalties, and temperature scaling is used to control the sharpness of the state assignmentdistributions. Training is performed by minimizing the negative log-likelihood of the observed sequences, with gradients propagated through both the transition and emission networks.

A. Transition Model

The Transition Model in the DifferentiableHMM is responsible for determining the probabilities of moving from one hidden state to another at each time step. Unlike traditional HMMs with static transition matrices, this model incorporates a history-aware, dynamic mechanism using neural networks to adapt transitions based on the patient's past observations and previous latent state.

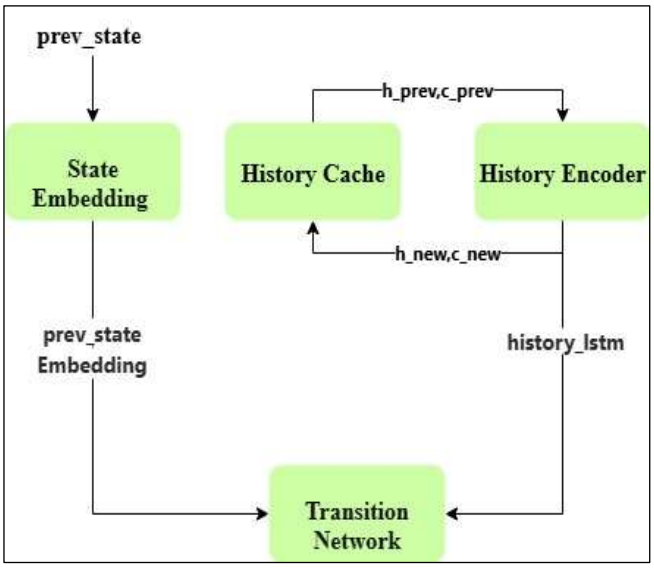


Figure 2: Transition Model Components.
Source: Authors, (2026).

Components of the Transition Model:

- **History Cache:**This component stores and retrieves the hidden and cell states of an LSTM for each patient across different batches. It ensures that the model can maintain a continuous memory of a patient's history, even when sequences are processed in batches.
- **History Encoder:**This is an LSTM network that processes the input observations (x) over time to generate a hidden state history (history_lstm). This history_lstm provides a rich, temporal context that informs the transition probabilities. The LSTM's states (hnew,cnew) are updated and stored in the history cache for subsequent time steps or future batches involving the same patient.
- **State Embedding:** Before being used in the transition calculation, the estimated latent state at the previous time step (prev_state_Embedding) is converted into a dense, continuous vector representation (embedding). This allows the discrete state to be effectively integrated into the neural network computations (see Figure 3).

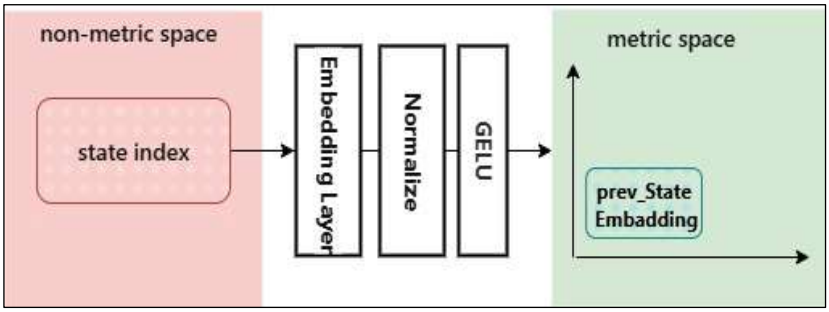


Figure 3. Hidden State Embedding.
Source: Authors, (2026).

- **Transition Network:** This is the core neural network that computes the raw logits for transitions between states. It takes the contextualized representation (from prev_state_Embedding and history_lstm) and outputs the log-probability of transitioning from one state to another. Finally, a softmax function is applied to the adjusted logits along the last dimension to ensure that the transition probabilities sum to 1 for each possible previous state.

B. Emission Model

The Emission model defines the conditional probability $p(x_t|z_t)$ of observing physiological measurements x_t given the current latent state z_t. To accommodate the complex, often multi-modal distributions of medical measurements, each latent state is associated with a Gaussian Mixture Model (GMM) in which each latent state z_t is associated with learnable component means and softplus-transformed scales ($means_{mixture}\backslash scales_{mixture}$) and a full covariance captured via a lower-triangular factor combined with a masking strategy to encode inter-feature dependencies. At each time step, a compact MixtureSelector network ingests sliding-window summaries of recent observations (x, x_{mean}, x_{std}) to produce dynamic mixture weights, providing a differentiable analogue of the expectation–maximization update in classical HMMs. During inference, component log-densities are aggregated via a numerically stable logsumexp and subjected to adaptive temperature scaling based on observed variance to sharpen the posterior over states. This design enables the model to flexibly capture complex, multi-modal medical measurement distributions and their temporal dynamics without assuming diagonal covariance or static mixture proportions.

III.2.3 Stepe3: training procedure

The training procedure for the DifferentiableHMM is meticulously designed to optimize its parameters for robust performance in medical time-series analysis. To this end, the model is trained by minimizing its negative log-likelihood (NLL) of the observed sequences, calculated based on the model's emission probabilities. Gradients for this loss are primarily propagated through the neural emission network, specifically involving the MixtureSelector that dynamically predicts Gaussian mixture weights based on input features. This primary loss objective is complemented by regularization terms for entropy and L2, serving to prevent overfitting and encourage diverse latent state utilization, with L2 regularization applying across all model parameters. A crucial initialization step involves setting the emission mixture means using K Means clustering on the training data, while emission scales are derived from the overall data standard deviation, enhancing training stability.

IV. EVALUATION

IV.1 DATASET AND EXPERIMENTAL SETUP

We evaluate DifferentiableHMM on the publicly available MIMIC-III intensive-care database [13], which includes de-identified records for over 40 000 ICU patients (2001–2012). Following prior work [14-16], we extract 100 patients with multiple ICU stays and six routinely monitored signals: DiastolicBP, HeartRate, Platelet, RespRate, Sodium, and SpO_2. Each patient trajectory is represented as a fixed-length sequence of 50 hourly measurements. We randomly stratify by patient into 80 % training and 20 % test sets. Hyperparameters (Table 2) are selected via greedy grid search on the validation split, optimizing the average MSE across prediction horizons k=1.10 We tune:

- Hidden units in the LSTM encoder and transition MLP
- Number of Gaussian mixtures per state
- Softmax temperature, entropy & L2 regularization
- Dropout rate, learning rate

To benchmark our approach, we compare against two strong baselines on held-out intensive-care time-series from the MIMIC-III dataset :

Classical HMM, trained via Baum–Welch with fixed transitions and Gaussian emissions.

Pure LSTM,LSTM networktrained to predict the next discrete state by cross-entropy.

Table 2: Key hyperparameter grid and chosen values.

Hyperparameter	Description	TypicalValues
n_components	Number of hidden states in the HMM	9
input_dim	Number of input features	6
hidden_dim	Hidden dimension for LSTM/history encoder and transition net	128
n_mixtures	Number of Gaussian mixtures per state	3
lr	Learning rate for optimizer	0.01
n_epochs	Number of training epochs	100
seq_length	Length of each time-series sequence	50

Source: Authors, (2026).

IV.2 QUANTITATIVE RESULTS

IV.2.1 Model Fit and Generalization

Table 3: Quantitative comparison of model fit and generalization metrics.

Model	LogLikelihood [17]	SeqNLL(perstep)	$R^2_{k=1}$	ECE
ClassicalHMM	8076.15	48.07	–	0.529
PureLSTM	–	–	0.530	0.473
DifferentiableHMM	4852.84	28.89	0.556	0.401

Source: Authors, (2026).

Sequence Negative Log Likelihood [18]: DifferentiableHMM's average per step NLL is 28.89 versus +48.07 for the classical HMM.

One Step PredictiveR2[19]: DifferentiableHMM (0.556) exceeds LSTM (0.530), demonstrating more accurate immediate forecasts.

ExpectedCalibration Error (ECE)[20]: DifferentiableHMM's lower ECE (0.401) shows its probability estimates are better calibrated.

IV.2.2 Multi Step Forecasting

To assess long-range dependency modeling, we forecast x_{t+k} for horizons k=1.10 and compute mean squared error (MSE).

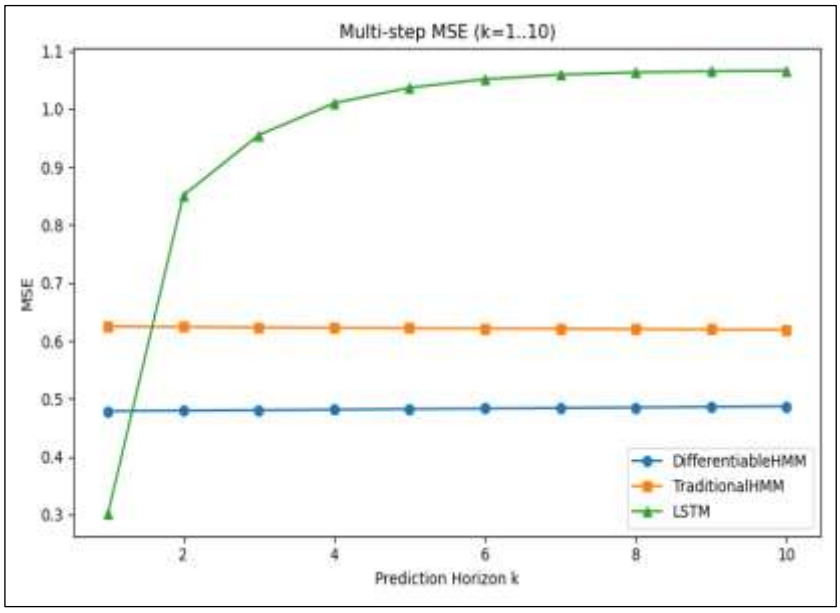


Figure 4: Multi-step MSE vs. prediction horizon.
Source: Authors, (2026).

In Figure 4 DifferentiableHMM (blue) maintains the lowest error across all, outperforming both Classical HMM (orange) and LSTM (green). DifferentiableHMM's flat MSE curve highlights its ability to leverage long histories via differentiable, history-conditioned transitions.

IV.3 INTERPRETABILITY AND CLINICAL ALIGNMENT

Table 4: Interpretability and Clinical Alignmentmetrics.

Model	State–Event MI	Cluster Purity	LatentEntropy	TransEntropy
Classical HMM	0.26	0.64	0.53	0.02
Pure LSTM	0.35	–	0.48	0.31
DifferentiableHMM	0.82	0.46	1.12	0.97

Source: Authors, (2026).

State–Event Mutual Information [21]: DifferentiableHMM's discrete states yield an MI of 0.82 with notated clinical events, far surpassing baselines.

Cluster Purity[22]: Although classical HMM exhibits higher raw clustering in feature space, DifferentiableHMM's superior alignment with events underscores its clinically meaningful state assignments.

Latent State Entropy[23]: At 1.12, DifferentiableHMM uses its nine-state capacity effectively (max2.20nats), avoiding collapse.

Transition Entropy[24]: DifferentiableHMM (0.97nats) balances state persistence and progression, whereas the classical HMM is near deterministic (0nats).

IV.4 ABLATION STUDY

We assess the impact of each core component by removing (i) the history encoder, (ii) differentiable EM, or (iii) Bayesian GMM initialization: To quantify the contribution of each core component of DifferentiableHMM, we performed three targeted ablations. For each variant, we reuse the emission parameters from the fully trained model to ensure a fair comparison. We then report the change (Δ) in three key metrics—Negative Log-Likelihood (NLL), one-step R^2, and State–Event Mutual Information (MI)—relative to the full DifferentiableHMM.

IV.4.1 Ablation Configurations

I. **Without History Encoder :**

Purpose: Evaluate the role of the history encoder (History Cache → LSTM → StateEmbedding → Transition Network) in capturing long-range patient context and informing transition probabilities.

Implementation: To disable the influence of the history encoder, a custom class, NoHistoryEncoder, was formulated. This class inherited from DifferentiableHMM, overridden to bypass the prev_state input, exclusively returning state probabilities derived from the emission model. The emission parameters (e.g., means_mixture, scales_mixture) were explicitly copied from the fully trained baseline model to ensure that any observed performance degradation was attributable solely to the absence of the history encoder and not to differing emission characteristics.

II. Without Differentiable EM (Mixture Selector)

Purpose: This experiment aimed to evaluate the contribution of the MixtureSelector neural network. This component is pivotal to the "differentiable EM" aspect, enabling the dynamic prediction and end-to-end learning of Gaussian mixture weights within the emission model via backpropagation.

Implementation: The NoDifferentiableEMHMM class was developed for this ablation uniform mixture weights were assigned. This effectively removed the adaptive, learned weighting mechanism, compelling each mixture component within a given state to contribute equally. As with the previous ablation, emission parameters were transferred from the trained full model.

III. Without GMM Initialization

Purpose: Probe the benefit of K-Means/GMM-based initialization of emission mixture means and scale parameters for stable convergence.
Implementation: the $(\text{means}_{\text{mixture}} \backslash \text{scales}_{\text{mixture}})$ parameter was populated with random normal values instead of the K-Means derived centres.

Table5: Ablation study results.

Variant	Δ NLL	Δ $R^2_{k=1}$	Δ State–Event MI
WithoutHistoryEncoder	+278.34	-0.0153	–0.0092
WithoutDifferentiable EM	+11.19	-0.0199	+0.0044
Without GMM Initialization	+272.47	-0.0105	–0.0001

Source: Authors, (2026).

History Encoder: Removing the LSTM-based history raises NLL by 278 and reduces $R^2_{k=1}$ by 0.015, confirming history conditioning is critical for both fit and predictive accuracy.

Differentiable EM: Disabling joint gradient flow increases NLL by 11 and slightly degrades R^2, illustrating the importance of end-to-end learning.

Bayesian GMM Initialization: Omitting informed emission priors also incurs a large NLL penalty (272), showing that robust state clustering at initialization is essential.

IV.5 DISCUSSION OF RESULTS:

DifferentiableHMM demonstrates superior performance over both classical HMMs and pure LSTMs across all evaluated dimensions—likelihood, forecasting accuracy, calibration, and interpretability. This consistent outperformance underscores the effectiveness of our core innovation: a differentiable, history-aware transition mechanism that captures complex clinical dynamics more faithfully than memoryless or black-box alternatives. Moreover, our ablation study reveals that removing any single component—whether history Encoder, end-to-end differentiability, or Bayesian-informed emission initialization—leads to measurable declines in model fit, predictive accuracy, and interpretability. These findings confirm that each element is indispensable to DifferentiableHMM's overall performance and validate the methodological contributions of this work.

V. CONCLUSIONS

We have presented DifferentiableHMM, a hybrid sequence-modeling framework that seamlessly integrates the interpretability of discrete Hidden Markov Models with the representational capacity of recurrent neural networks. By conditioning transition probabilities on an LSTM-encoded history, employing a fully differentiable EM algorithm for joint optimization, and initializing emission mixtures via Bayesian Gaussian mixture clustering, our approach addresses the fundamental limitations of classical HMMs—namely, their memorylessness and rigid transition structure—as well as the lack of transparent latent states in purely neural models. Comprehensive experiments on the MIMIC-III intensive-care dataset demonstrate that DifferentiableHMM substantially outperforms both a standard HMM and a pure LSTM baseline. In particular, our method achieves markedly lower sequence negative log-likelihood, higher one-step predictive R2, improved calibration, and considerably stronger alignment between inferred states and clinically annotated events. Moreover, the learned transition dynamics exhibit a realistic balance between state persistence and progression. An extensive ablation study confirms that each architectural component contributes critically to these gains. Looking forward, we plan to extend DifferentiableHMM in several directions. First, we will incorporate patient-level covariates (such as demographics and treatment interventions) directly into the transition network to enable personalized state dynamics. Second, deployment in real-time monitoring scenarios could facilitate early detection of critical events by tracking interpretable state transitions online. Finally, we will integrate formal differential privacy guarantees into the learning process, thereby enabling privacy-preserving state inference and synthetic data generation with provable protection of sensitive patient information.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Arbia Boudaoud, Salheddine Kabou.
Methodology: Arbia Boudaoud, Salheddine Kabou.
Investigation: Arbia Boudaoud, Salheddine Kabou.
Discussion of results: Arbia Boudaoud, Salheddine Kabou.
Writing – Original Draft: Arbia Boudaoud, Salheddine Kabou.
Writing – Review and Editing: Arbia Boudaoud, Salheddine Kabou.
Resources: Arbia Boudaoud, Salheddine Kabou.
Supervision: Arbia Boudaoud, Salheddine Kabou.
Approval of the final text: Arbia Boudaoud, Salheddine Kabou.

VII. REFERENCES

[1] S. Kabou, S. M. Benslimane, and M. Mosteghanemi,"A survey on privacy preserving dynamic data publishing", In Research Anthology on Privatizing and Securing Data (pp. 1635-1657). IGI Global, 2021.

[2] H. Kuehne, A. Richard, and J. Gall, "A hybrid RNN-HMM approach for weakly supervised temporal action segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 4, pp. 765–779, 2018. DOI: 10.1109/TPAMI.2018.2884469.

[3] Y. Sun, L. T. Bosch, and L. Boves, "Hybrid HMM/BLSTM-RNN for robust speech recognition," in Text, Speech and Dialogue: 13th Int. Conf., TSD 2010, Brno, Czech Republic, Sept. 6–10, 2010, Proc. Berlin, Germany: Springer, 2010, pp. 400–407. DOI: 10.1007/978-3-642-15760-8_51

[4] H. Dai, B. Dai, Y. M. Zhang, S. Li, and L. Song, "Recurrent Hidden Markov Model," in Proc. Int. Conf. Learn. Represent. (ICLR), Feb. 2017.

[5] J. Yu, L. Xie, X. Xiao, and E. S. Chng, "A hybrid neural network hidden Markov model approach for automatic story segmentation," J. AmbientIntell. HumanizedComput., vol. 8, pp. 925–936, 2017. DOI: 10.1007/s12652-017-0501-9

[6] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks conceptstohybridnn-hmmmodelforspeech recognition," in 2012 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP). IEEE, Mar. 2012, pp. 4277–4280. DOI: 10.1109/ICASSP.2012.6288864

[7] H. Zen, "Acoustic modeling in statistical parametric speech synthesis—from hmm to LSTM-RNN," in Proc. MLSP, 2015.

[8] S. Mehta, Székely, J. Beskow, and G. E. Henter, "Neural hmms are all you need (for high-quality attention-free TTS)," in ICASSP 2022– IEEE Int. Conf. Acoust., Speech Signal Process. IEEE, May 2022, pp. 7457–7461. DOI: 10.1109/ICASSP43922.2022.9746686.

[9] Y. Liu, Z. Zhang, and S. Qin, "NeuralHMM: A deepmarkov network for health risk prediction using electronic health records," in 2023 Int. Joint Conf. Neural Networks (IJCNN). IEEE, Jun. 2023, pp. 1–8. DOI: 10.1109/IJCNN54540.2023.10191594.

[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, no. 2, pp. 257–286, Feb. 2002. DOI: 10.1109/5.18626.

[11] C. B. Do and S. Batzoglou, "What is the expectation maximization algorithm?" Nat. Biotechnol., vol. 26, no. 8, pp. 897–899, Aug. 2008.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997. DOI: 10.1162/neco.1997.9.8.1735

[13] A. E. Johnson and et al., "Mimic-iii, a freely accessible critical care database," Sci. Data, vol. 3, no. 1, pp. 1–9, 2016. DOI: 10.1038/sdata.2016.35

[14] F. Faisal, N. Mohammed, C. K. Leung, and Y. Wang, "Generating privacy-preserving synthetic medical data," in Proc. IEEE 9th Int. Conf. Data Sci. Adv. Analytics (DSAA). IEEE, 2022, pp. 1–10, DOI: 10.1109/DSAA54385.2022.10032429

[15] H. Karami, M. A. Hartley, D. Atienza, and A. Ionescu, "Timehr: Image-based time series generation for electronic health records," IEEE J. Biomed. Health Inform., 2025. DOI: 10.1109/JBHI.2025.3577328

[16] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional GANs," 2017, arXiv:1706.02633.

[17] Y. Takakura, J. Mori, and H. Kobayashi, “A log-likelihood-based evaluation metric for the reproducibility and simplicity of logistics graphs,” IFAC-PapersOnLine, vol. 56, no. 2, pp. 6174–6180, 2023. DOI: 10.1016/j.ifacol.2023.10.731

[18] L. C. Barbosa, “Maximum likelihood sequence estimators: A geometric view,” IEEE Trans. Inf. Theory, vol. 35, no. 2, pp. 419–427, Mar. 1989. DOI: 10.1109/18.32136

[19] Y. Babich, L. Hlazunova, T. Kalinina, and Y. Petrovych, “R2 metric dynamics for k-nearest neighbors regression model trained on series of different sizes,” Infocommun. Technol. Electron. Eng., no. 4, pp. 10–18, 2024.

[20] N. Posocco and A. Bonnefoy, “Estimating expected calibration errors,” in Artif. Neural Nets. Mach. Learn.-ICANN 2021: 30th Int. Conf. Artif. Neural Nets., Bratislava,Slovakia, Sept. 14–17, 2021, Proc., Part IV. Springer,2021, pp. 139–150. DOI: 10.1007/978-3-030-86380-7_12

[21] A. M. Fraser, “Using mutual information to estimate metric entropy,” in Dimensions and Entropies in Chaotic Systems: Quantification of Complex Behavior. Berlin, Germany: Springer, 1986, pp. 82–91.

[22] J. L. V. Noguera, E. L. Pezoa, G. Espínola, and S. A. Grillo, “A symmetric purity measure for clustering comparison,” Ann. Data Sci., pp. 1–15, 2025. DOI: 10.1007/s40745-025-00594-z

[23] M. Kocaoglu, S. Shakkottai, A. G. Dimakis, C. Caramanis, and S. Vishwanath, “Entropic latent variable discovery,” Stat., vol. 1050, p. 26, 2018.

[24] T. M. Cover, Elements of Information Theory. New York, NY, USA: John Wiley & Sons, 1999 DOI:10.1002/047174882X

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 133-146. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2805

RESEARCH ARTICLE OPEN ACCESS

ENHANCING RENEWABLE ENERGY RELIABILITY: A STUDY ON HYBRID SOLAR-WIND SYSTEMS FOR EFFICIENT POWER GENERATION

Omar Gheyath Ibrahim[1], Mohammed Tariq Majeed[2]
Ghufran Saad Mohammed[3], Israa Khalouq Saaed[4]

[1]Dept. Of Medical Equipment Techniques, Polytechnic College Mosul, Northern Technical University, Mosul, Iraq.
[2]Dept. Of Medical Equipment Techniques, Polytechnic College, Northern Technical University, Mosul, Iraq.
[3]The General Company for Northern Electricity Distribution, Ministry of Electricity of Iraq.
[4]Dept of Mechatronic, Polytechnic College, Northern Technical University, Mosul, Iraq.

[1]https://orcid.org/0009-0008-0917-2454, [2]https://orcid.org/0000-0001-6885-8597
[3]https://orcid.org/0009-0004-6298-8146, [4]https://orcid.org/0009-0004-5236-7370

Email: mti.lec17.omar@ntu.edu.iq, aldabgh68@ntu.edu.iq, ghufran.22enp41@student.uomosul.edu.iq, mti.lec140.israa@ntu.edu.iq

ARTICLE INFO

Article History
Received: October 10, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
HOMMER, photovoltaic (PV), Global Horizontal Irradiance (GHI),

ABSTRACT

Hybrid solar–wind plants offer an effective resolution of the real problem of intermittency, ensuring steady electricity supply and minimizing the need for technology of storage. Although they cost more upfront, they provide savings over time, aid the environment and advance decarbonization targets. HOMER simulation for Basra also showed the predominance of solar power where the total annual capacity was 10.54 kWh/m², while wind provided 1.21 kWh, confirming the efficiency of hybrid system integration for Basra as a typical example. Worldwide experience, including Germany, China and the U.S., have also shown that systems such as hybrids are more reliable and cost effective than single platforms. These results emphasize how important the phasing control will be in the upcoming energy transitions, and the importance of advances in storage technologies, AI-based control, and the role of legislation in making these precedents.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Amid the global concern regarding climate change and the dissipation of fossil fuel supplies, the need for sustainable, and renewable sources of energy has never been so pressing. Traditional, fossil fuel-based energy systems cause the release of greenhouse gases and pollution, and drive a spiraling global disaster 1. In this field, renewable energies, among these the solar, wind and hydro power sources, are increasingly the focus of interest in order to obtain clean, yet abundant and sustainable electricity. Among these, solar and wind energy are two of the most promising and widely-used technologies for the massive energy generation [3]. However, both energy sources suffer from intermittency, low efficiency and reliability, which makes it difficult to satisfy the whole global energy need when vertically used [4]. The solar energy, which is obtain through photovoltaic (PV) cells, entirely relies upon sunlight that varies greatly according to weather conditions, time frame and geographical locations [5].

Likewise, wind energy produced by wind turbines varies as a result of change in wind speed and it is affected by seasonal variations and local weather condition. However, these natural intermittencies are causing problems with maintaining a consistent and dependable electricity supply, at least in places where the grid has high dependence on renewable sources[6]. To counter these challenges, hybrid energy systems involving solar and wind power have emerged as an attractive alternative to mitigate intermittency challenges while enhancing the overall reliability and efficiency of renewable energy generation [7]. A combination of wind and solar in a single system can use the complementary characteristics on these two sources. Solar power works best during the day, while wind energy can be more prevalent at night or when it is not enough light [8]. The collaboration between both systems can help make sure that the power supply is more dependable and stable all day throughout the year.

This renders hybrid systems an excellent option for producing electricity in a way that is good for the environment [9]. Hybrid energy systems had several important advantages over traditional energy generation systems that operate on its own. First, they have a greater capacity factor, which is a ratio of actual energy produced to the greatest possible energy output. It's due to their ability of making power from both solar and wind sources [3]. So, the system works better overall and doesn't need as many backup power sources, such fossil fuel-based power plants, which are employed when renewable energy supply is low [10]. Combining wind and solar electricity can also lower the cost of storing and sending power because it means you don't need massive storage facilities to handle times when one of the sources is making a lot of power [11]. The creation and use of hybrid energy systems also fits with the global goal of moving to a low-carbon economy. A lot of countries are now establishing high goals for cutting greenhouse gas emissions and raising the amount of renewable energy in their energy mix [12]. Hybrid systems can help fulfil these goals by offering a more steady and reliable source of renewable energy that can help integrate renewables into the grid as a whole [13].

In this context, renewable energy sources such as solar, wind, and hydropower have gained significant attention due to their potential to provide clean, abundant, and sustainable electricity. Figure (1) shows how the quantity of renewable energy capacity added will change between 2019 and 2024. It also shows that wind energy will grow faster than solar energy. Wind energy won by a long shot in 2019, with 119.1 GW, while solar energy only added 59.4 GW. This pattern continued in 2020, when wind power rose to 153.2 GW and solar power grew to 107.8 GW. In the next few years, wind energy increased a lot faster. It grew to 232 GW in 2022 and then to 426.3 GW in 2023, which was faster than the constant rise of solar energy. Solar energy started getting better, but it was still slower than other kinds of energy, especially between 2021 and 2022. By 2024, wind energy had risen to an incredible 553 GW, but solar energy had only expanded to 120 GW. Because of new technology like offshore wind farms and cheaper costs, this research suggests that wind energy will become the dominant source of renewable energy capacity. Solar energy, on the other hand, has been accepted more slowly. This is partly because it doesn't always work, it costs more to install at initially, and there are issues with how to use the land. Both of these sources of energy are still very vital for the global endeavor to make the future low-carbon. Wind energy is becoming the main focus of big renewable energy projects, while solar energy is still a big part in locations with a lot of sun [14].

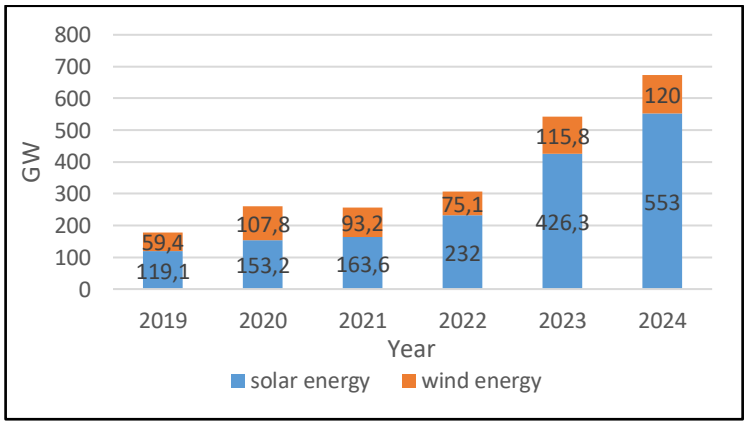


Figure 1: Renewable Energy Capacity Additions by Technology (2019-2024): Solar vs. Wind.
Source: Authors, (2026).

II. HYBRID SOLAR-WIND SYSTEMS:

Hybrid energy systems use both solar and wind power to make it easier and more dependable to get electricity from renewable sources. Figure (2) shows that these systems try to solve the problems of intermittency and variability that each source has when used alone by combining both technologies [15].

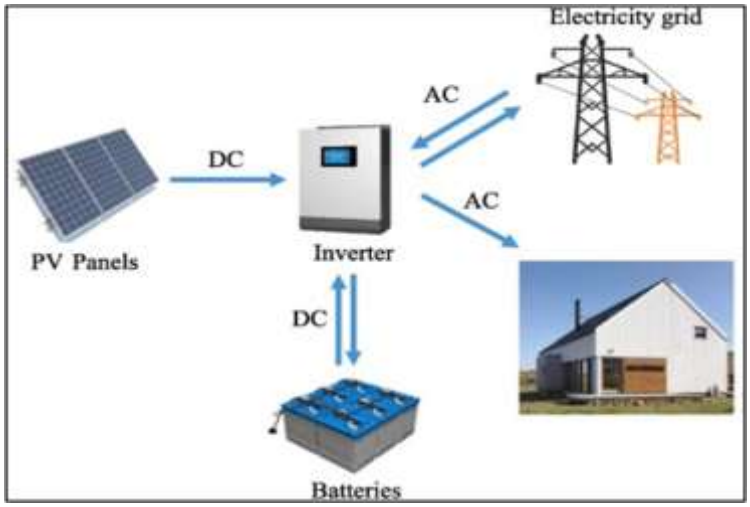


Figure 2: Hybrid Energy System.
Source: Authors, (2026).

II.1 KEY FEATURES OF SOLAR-WIND HYBRID SYSTEMS:

- Complementary Generation:

Solar and wind power operate well together. During the day, when there is sunlight, solar energy is usually abundant. Wind energy, on the other hand, is generally more useful at night or on overcast days when solar energy is low. [16].

- Improved Efficiency and Reliability:

Combining both sources makes energy generation more reliable resulting in fewer occasions when energy isn't available and a more stable power supply overall. This is especially helpful in areas where the weather changes a lot or the sun shines differently. [17].

- Reduced Need for Energy Storage:

The integration of solar and wind energy reduces the requirement for large-scale energy storage systems. The hybrid system can alleviate periods of underperformance from a single source, reducing dependence on storage devices and lowering costs.[18].

- Geographical Versatility:

These systems can be deployed in regions where either solar or wind alone might not be sufficient. Areas with both wind and solar potential can benefit greatly from hybrid systems, providing a more balanced and consistent energy output [19].

II.2 ADVANTAGES OF SOLAR-WIND HYBRID SYSTEMS

- Increased capacity factor: The system generates electricity more reliably, even during periods when one source is not performing.
- Lower dependence on backup generation: Reduces reliance on fossil-fuel-based power plants or expensive storage systems [20].
- Environmental benefits: Decreases greenhouse gas emissions and supports a sustainable energy transition.
- Scalability: Hybrid systems can be expanded or adjusted according to the energy demands and geographical conditions [21].

III COMPARISON BETWEEN STANDALONE AND HYBRID SYSTEMS:

Standalone systems are renewable energy systems that only use one source of electricity, such solar or wind power. These systems can be helpful in some situations, but they often have problems with variability and intermittency. Based on a lot of different things, Table 1 shows how solar-wind hybrid systems compare to standalone systems [22][23][24]:

Table 1: Comparison of Solar-Wind Hybrid Systems vs. Standalone Renewable Energy Systems.

Feature	Solar-Wind Hybrid Systems	Standalone Systems (Solar or Wind Alone)
Energy Generation Reliability	High reliability due to complementary nature of solar and wind; stable power output across the day and seasons	Lower reliability; dependent on the availability on one source, leading to gaps in generation
Intermittency	Reduced intermittency, as wind can generate power when solar is not available and vice versa	High intermittency; solar energy is unavailable at night and on cloudy day; wind energy depends on wind conditions
Efficiency	Higher overall efficiency, both sources contribute to consistent power generation	Lower efficiency; relies solely on on ernergy source, leading to periods od low generation
Cost of Installation	Higher initial costs du to the complexity of integrating both systems	Lower initial costs for installation compared to hybrid systems, as only one energy source is used
Energy Storage Needs	Reduced need for large scale energy storage systems; excess energy generation can be stored in batteries or used in real-tine	Higher energy storage requirments, especially during periods of low generation, requiring expansive battries or backup systems
Space Requirements	Requires more space for both solar panels and wind turbines, but offers better land utilization when integrated	Less space required compared to hybrid systems but still needs significant space for either solar panels or wind turbine
Operational Complexity	More complex to manage and operate; requires advanced control systems to optimize energy production and integration	Simple operation but may require backup or grid connection to maintain reliability
Scalability	Scalable; can be easily expand by adding more solar panels or wind turbines as demand increases	Limited scalability as each system relies on a single source of energy. Expanding capacity requires increasing the size of the same source
Geographical Suitability	Suitable for areas with variable weather conditions and diverse climates where both solar and wind potential exists	Best suited for areas where either solar or wind is highly predictable and abundant

Source: Authors, (2026).

IV. PERFORMANCE AND ENVIRONMENTAL/ECONOMIC ASPECTS:

Sunlight, wind, temperature, and seasons substantially affect hybrid systems. Where you reside affects solar and wind resource availability, which affects energy generating reliability and efficiency. Figure (3) shows that hybrid energy systems using solar and wind energy function differently depending on location.

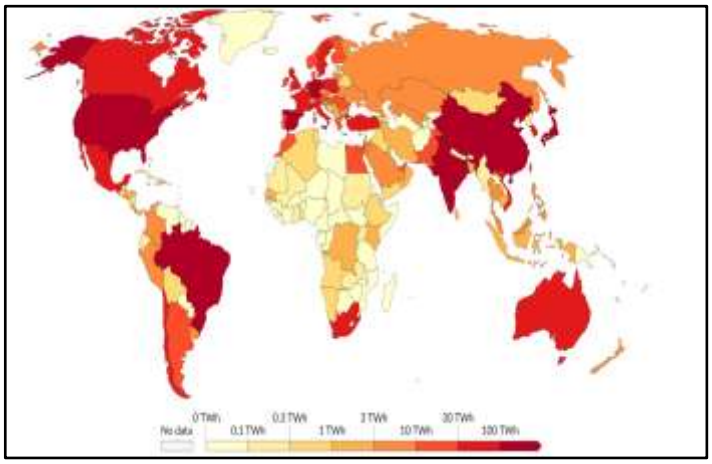


Figure 3: Solar and wind power generation, 2024.
Source: [25].

In deserts with a lot of sunshine and little wind, solar panels function more effectively than wind turbines. Wind turbines produce more electricity as the wind is steady, like on the coast or in mountainous areas, solar backups is used. Hybrid systems work better in these areas because they use the light during the day and the wind at night or on cloudy days. It makes the energy supply more stable [26]. Some island nations and remote areas have both solar and wind potential, and hybrid systems can be very useful there. They make sure that energy is always available and that you do not have to rely on outside grids. Hybrid systems may not work as well or cost as much in places where one source is more common. The performance and productivity of hybrid systems depend on environmental factors. In places with a lot of both wind and solar energy, like the coast or the desert [27], hybrid systems work better than solar or wind systems on their own. In desert areas, for instance, a solar-wind hybrid system mostly uses solar power. Wind turbines, on the other hand, make most of the power in places with a lot of wind along the coast [28]. Hybrid systems make sure that the amount of energy produced stays the same by making up for low performance from one source with the other. Hybrid systems make things more reliable in places where the seasons change a lot by using wind energy when the sun is not shining as much [29].

IV.1 SEVERAL WEATHER AND CLIMATE FACTORS PLAY A CRUCIAL ROLE IN DETERMINING THE PERFORMANCE OF HYBRID ENERGY SYSTEMS. THESE INCLUDE:

- Solar Radiation and Sunshine Hours:

The amount and length of sunlight hours are very important for solar power generation. These change with the seasons and latitude. For example, areas near the equator have longer days all year, which means they get more solar energy. Conversely, areas at higher latitudes, like those near the poles, have shorter days in the winter, which can make solar panels less effective [29].

- Wind Speed and Patterns:

The strength and steadiness of the winds determine how much wind energy can be made. The area's geography, its height, and how close it is to the coast or mountains are some of these things. Wind turbines work better on the coast and in the mountains because the winds are stronger and more steady there. The wind also changes with the seasons, and in some places, these changes can change how much power turbines make. [30].

- Temperature and Humidity:

The efficiency of both solar panels and wind turbines depends a lot on the temperature. In general, solar panels work better when it is cooler outside. This is because extreme heat can make the cells less efficient by making them more resistant. Wind turbines, on the other hand, are affected by the density of the air, which goes down as the temperature goes up. This makes them less efficient at making energy. Humidity is also important because high levels of moisture in the air can block solar radiation, which makes solar panels work less well. [31].

- Cloud Cover and Precipitation:

Cloud cover and rain can block sunlight from reaching solar panels, especially in places where the sky is often cloudy [32]. Rain or snow can also make wind turbines work less well by making ice build up or changing the speed of the wind. Hybrid systems, on the other hand, are better at handling these changes because wind turbines can keep working even when solar generation drops. [33].

- Environmental Obstacles:

The shape of the land and the presence of natural barriers like mountains, forests, or buildings can change the direction of the wind, creating wind shadows or turbulence that can make wind turbines less efficient. The orientation of the panels, shade from nearby objects, and local microclimates can all affect how well solar systems work. [34].

IV.2 ENVIRONMENTAL AND ECONOMIC CHALLENGES OF HYBRID ENERGY SYSTEMS:

Hybrid energy systems, like solar-wind combinations, have many benefits, such as reducing our dependence on fossil fuels and lowering greenhouse gas emissions. However, they also have some environmental problems [35]. Some of the main environmental issues that come up with renewable energy systems are:

- Land Use and Habitat Disruption:

Large space is needed for Large-scale hybrid system for Bothe solar panals and wind turbine. This can hurt habitats, especially in places where the environment is very sensitive. For example, wind farms can change the routes that birds take when they migrate, and big solar farms may have to cut down trees and other plants. But you can lessen these effects by picking the right place and doing environmental impact assessments. [36].

- Resource Extraction for Manufacturing:

To make solar panels and wind turbines, you need to get raw materials like copper, silicon, and rare earth metals. When these materials are mined, they can harm the environment, pollute water, and release more carbon into the air. To lessen the bad effects of these materials, it is very important to recycle them and use sustainable mining methods to lower their environmental impact. [37].

- End-of-Life Disposal and Recycling:

Another environmental problem is getting rid of old solar panels and wind turbine blades. Solar panels usually last 25 to 30 years [38], and wind turbines usually last 20 to 30 years [39]. When these parts reach the end of their useful life, they need to be recycled or thrown away in a safe way. If not done right, this can be expensive and bad for the environment.

- Noise Pollution and Aesthetic Impact:

People often complain about how noisy and ugly wind turbines are [40]. The noise from the turbines can bother animals and people who live nearby, and big wind farms can change the way rural or coastal areas look. These problems can be lessened by putting the turbines in the right places and making them quieter and more visually appealing[41].

IV.3 ECONOMIC FEASIBILITY ANALYSIS:

Initial Costs, Long-Term Benefits, and the Cost of Energy Generated Hybrid energy systems have high initial costs and long-term economic benefits, which is a problem from an economic point of view. The cost of installing, maintaining, and running these systems is one of the things that affects how economically feasible they are.

- Initial Costs:

Setting up a hybrid power system typically costs more upfront than implementing in a solar or wind system alone. The installation may also need improvements to the grid and energy storage systems, which will make the initial costs even higher. [42].

- Long-Term Benefits:

Over time, hybrid systems can save a lot of money compared to traditional systems that use fossil fuels. The only ongoing costs for hybrid systems are maintenance, since they do not use any fuel. Also, they can make electricity for 20 to 30 years or more, which means that after the initial investment, they provide a stable and predictable power supply at a lower cost. [44].

- Cost of Energy Production:

The Levelized Cost of Energy (LCOE) is a common way to compare the costs of making electricity from different types of energy. Because of the cost of installation, hybrid systems usually have a higher LCOE in the first few years. But the cost usually goes down over time because the system makes energy more efficiently and the need for outside power sources goes down [45]. The LCOE for hybrid systems is usually higher than that of traditional fossil fuel-based generation. However, when you look at long-term savings, environmental benefits, and lower carbon footprints, it is competitive with or even lower than many other renewable energy systems. [46].

V. GOVERNMENT POLICIES AND GLOBAL TRENDS

Policies made by the government are very important for the growth and development of hybrid energy systems. By using different incentives, rules, and programs [47], policymakers can change how renewable energy technologies are used and how they grow. There are many ways that governments can help hybrid systems, such as:

- Money: Governments often give people money to help them pay for the costs of putting in renewable energy systems. This could be in the form of grants, tax breaks, or government money. These incentives can help homeowners and businesses buy hybrid energy systems [48].

- Support for Research and Development (R&D): Governments can pay for R&D projects that make hybrid energy systems work better and more efficiently, especially when it comes to cutting costs, storing energy, and connecting to the grid. Governments can help lower the overall cost of renewable energy and make hybrid systems more likely to be used by a lot of people by encouraging new ideas [49].
- Regulatory Support and Grid Access: Governments can make it easier for hybrid systems to connect to the grid by making rules. These could be rules about who can use the grid, rules about net metering, and rewards for utilities that use clean energy sources. Governments can also make long-term plans for energy that put clean energy and renewable integration at the top of the list. This will help keep the market for hybrid systems stable [50].
- Public Awareness and Education: Governments can also help by teaching people about the benefits of renewable energy and hybrid systems in general. People are more likely to accept and use new ideas if they learn about them in public and are given reasons to get involved in the private sector [51]. The table (2) below compares important countries and emerging markets based on their government policies, economic viability, and success in developing renewable energy, with a focus on solar-wind hybrid systems. The comparison shows how different factors affect the use and growth of hybrid energy systems in different areas. It looks at things like how far we have come in combining different types of renewable energy, how well the rules work, and how much money they make.

Table 2: Comparison of Government Policies and Renewable Energy Development in Key Countries and Emerging Markets for Hybrid Systems.

Country	Government Support	Renewable Energy Capacity	Economic Feasibility	Challenges	Notable Developments	Ref.
Germany	Strong financial incentives, grid integration policies, and ambitious renewable energy targets	One of the largest solar and wind markets in Europe.	High initial costs, but strong long-term returns.	Land use conflicts, grid integration challenges.	Leading in energy transition (Energiewende), a robust market for hybrid systems.	[52] [53]
United States	Federal tax incentives, state-level renewable energy programs, and R&D funding.	Significant growth in wind and solar, with some hybrid projects.	Moderate upfront costs, but substantial long-term savings.	Varies by state; inconsistent policies.	Increased hybrid system adoption in rural and remote areas.	[54]
China	Strong government subsidies, investments in R&D, and national renewable energy targets.	World's largest producer of solar energy, fast-growing wind energy market.	Lower LCOE due to government support and scale of production.	Environmental impacts of large-scale wind and solar farms.	Leading global player in solar and wind, with growing interest in hybrid systems.	[55]
India	Financial incentives for renewable projects, state-level solar and wind policies.	Rapid growth in solar and wind energy sectors, potential for hybrid systems.	High initial costs but government support for financing.	Grid reliability issues and infrastructure gaps.	Hybrid solar-wind projects in rural areas and remote locations.	[56]
Brazil	Government incentives for wind and solar, favorable policies for hybrid systems.	Expanding wind and solar capacity, especially in the Northeast.	Competitive costs due to abundant natural resources.	Regulatory hurdles and land use concerns.	Emerging market for hybrid systems, particularly in remote areas.	[57]
South Africa	Feed-in tariffs, tax incentives for renewable energy projects.	Significant potential for both wind and solar energy.	High upfront costs but government programs to support financing.	Grid integration and high capital costs.	Increasing investment in renewable energy and hybrid systems.	[58]

Source: Authors, (2026).

The world needs more energy, that's why many countries are looking into other, more sustainable ways to get it. Figure(4) shows that looking at energy use trends from 2015 to 2024 can tell us a lot about how hybrid energy systems are used, especially in countries where energy needs are growing quickly.By2024, China's energy use is expected to reach its highest level, which is almost 1,800 TWh.This is because the number of people and businesses is growing. In order to cut down on its reliance on fossil fuels, China will need to use more types of energy, like solar-wind hybrid systems. The U.S. rises more slowly, reaching about 800 TWh by 2024. This is because energy efficiency has gotten better. But hybrid systems are still good for the environment and the economy in the long run. Countries such as India, Germany, Brazil, and South Africa are experiencing moderate or sluggish growth. India's steady rise in energy use makes hybrid systems more useful.

The other countries, on the other hand, are growing more slowly, which shows that they need targeted and cost-effective renewable energy solutions Factories are using more and more energy, so China and the U.S. are ready to use hybrid systems on a large scale. Emerging markets might have trouble with the initial costs, but if their governments make good policies and offer financial incentives, they can adopt these technologies at their own pace. In countries where growth is slow,campaigns to raise public awareness and teach people can help them adopt even faster. Hybrid systems are the best way to keep energy prices stable over time, cut down on pollution, and make sure that energy will still be available in the future.New and rapidly growing markets need certain rules and incentives to make sure that they are widely used [59].

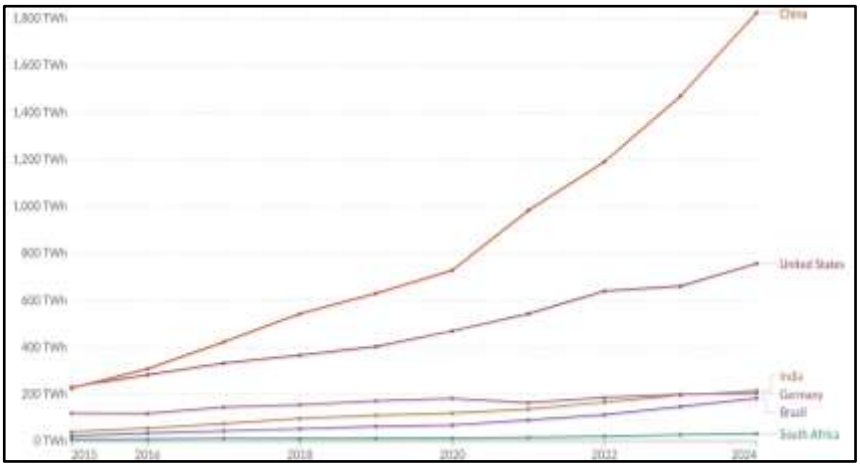


Figure 4: Energy Consumption Trends (2015-2024) for Major Economies.
Source: Authors, (2026).

VI. PRACTICAL APPLICATIONS AND FUTURE EXPANSION OF HYBRID SYSTEMS:

Solar-wind hybrid energy systems have been successful in both small villages and large power plants, generating electricity in areas where conventional energy infrastructure may be ineffective. It is too expensive to extend the regular electricity grid to these systems, which are useful in rural areas, islands, and remote communities [60]. These hybrid systems can give you a steady and reliable supply of electricity by using both solar and wind power. This can make life better and help the economy grow. In places where there is no grid access, hybrid systems can be used instead of expensive and bad for the environment diesel generators for backup power [61]. Hybrid solar-wind systems are better for the environment, last longer, and mean we do not have to buy fossil fuels from other countries as much. Energy storage systems make the system stronger by making sure that power is always there, even when it is dark or there is not much wind. Hybrid systems can save money over time, even in low-income areas, because they use less fuel and need less maintenance [62].

Hybrid systems give big power plants renewable energy that is more stable and flexible, which helps keep the grid stable. Hybrid systems make energy sources less volatile by using both solar and wind power. This is very important for getting more renewable energy into national systems because solar and wind power are not always reliable [63]. In large grids, hybrid systems can reduce the need for fossil-fuel power plants, which helps to create a more sustainable energy system. Hybrid systems also help electrical grids use more energy from sources that do not run out. Many countries are cutting their carbon footprints to meet international climate goals [64] as the demand for cleaner energy grows. It is easier and more reliable to add renewable energy to the grid when there are hybrid systems that use two sources of energy. This means that you can add more renewable energy to the grid without making the system less stable. If one power source stops working, the other can take over and keep the flow of energy going. This capacity makes it easier to handle at scale by reducing problems with intermittency [65]. This helps grids move toward more renewable energy.

Hybrid energy systems of the future will get a lot better, especially when it comes to how they store and control energy. We will need better ways to store renewable energy as we make more of it so that it can handle changes in energy sources. In hybrid systems, lithium-ion batteries store energy when it is at its highest and release it when it is at its lowest. More and more people need storage systems that work better, cost less, and last longer. People are looking into solid-state batteries, flow batteries, and compressed air energy storage (CAES) as ways to improve storage systems. These changes should make hybrid systems work better by making them last longer and store more energy [66]. Smart grid technology, better control systems, and better storage are the best ways to make hybrid energy systems better. To efficiently share power between solar, wind, and storage, hybrid systems need advanced power management [67]. AI and machine learning can improve control systems, which means that hybrid systems can see changes in how much energy is being made and used and change their behavior in real time.

These smart systems decide on their own how to store energy, feed it into the grid, and use it locally, which makes energy use and waste less of a problem [68]. In addition, demand-side management (DSM) systems that let people change how much electricity they use based on how much it costs and how much is available will be more important in hybrid energy systems. When you combine DSM with hybrid power systems, you can better match energy use with patterns of renewable generation. This makes the whole system work better. Solar wind hybrid energy systems are a new way to deal with the problems that come up when renewable energy is not always the same and changes [69]. These systems make electricity more reliably by using both solar and wind power. This is especially helpful in areas where neither is enough. Hybrid systems are more dependable because they can store energy, which means they can keep the power on even when generation is low. In remote areas and big power plants, hybrid systems are a long-lasting and cost-effective alternative to fossil fuels [70]. Hybrid systems use less fossil fuels and do less harm to the environment than traditional ways of making energy. Problems include using land, getting materials to make things, and getting rid of old parts. New designs, recycling, and managing resources in a way that does not hurt the environment can help make these big problems smaller [71]. It costs more to install hybrid systems, but they are a good investment because they save money on fuel, lower operating costs, and help the environment in the long run [72]. Even in poor countries, hybrid systems are becoming more affordable thanks to government subsidies, incentives, and research funding [73].

VII. CASE STUDY (IRAQ-BASRA): SIMULATION USING HOMER

The main focus of this study is the global trends, advantages, and issues associated with solar-wind hybrid systems. But HOMER software was used to run a localized simulation to show how these systems could work in the real world. We picked Basra, Iraq, as a good example because it gets a lot of sun and has a moderate amount of wind, which fits with what we talked about in the global context. This simulation does not narrow the study's focus; rather, it demonstrates the practicality, cost-effectiveness, and utility of hybrid systems in real-world contexts. A case study simulation was conducted using HOMER Pro software to validate the theoretical foundations and evaluate practical applicability, as detailed in the following section.

VII.1 ANALYSIS OF SOLAR GLOBAL HORIZONTAL IRRADIANCE (GHI) AND CLEANNESS INDEX DATA

This study analyzes the Monthly Average Solar Global Horizontal Irradiance (GHI) and Cleanness Index data, as shown in Figure (5), to understand the solar energy potential and atmospheric conditions impacting solar power generation. The following research provides information on monthly and annual solar radiation, trends in the cleanness index, and the potential for solar panels to generate power.

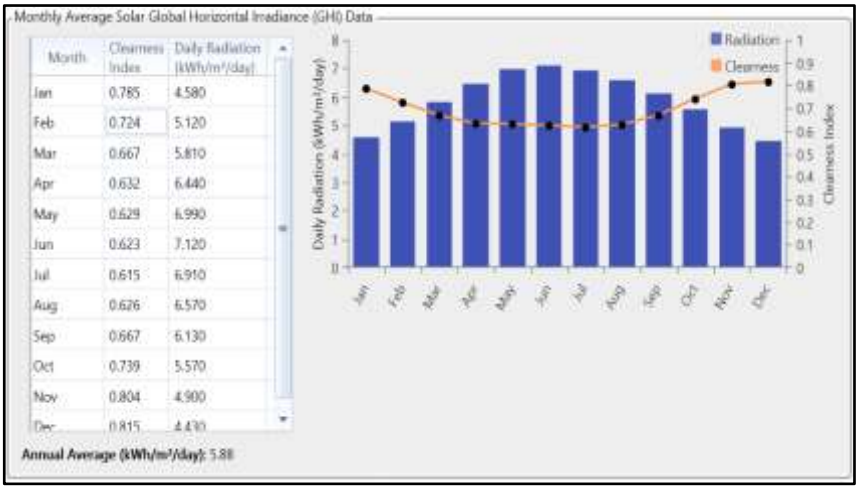


Figure 5: Monthly Distribution of Solar Irradiance and Clearness Index from HOMER Simulation (Basra, Iraq).
Source: Authors, (2026).

VII.1.1 Annual Average Solar Radiation

Over the course of the year, 5.88 kWh/m²/day is the average amount of solar energy received. The basis for assessing the potential for solar energy in various seasons is this figure.

VII.1.2 Monthly Breakdown of Solar Radiation

To illustrate how solar energy potential fluctuates throughout the year, Table 3 presents a monthly breakdown of solar radiation (in kWh/m²/day) and related energy output (in kWh). Understanding the monthly effectiveness of solar power generation requires an understanding of these figures.

Table 3: Monthly Breakdown of Solar Radiation and Energy Production.

Month	Solar Radiation (kWh/m²/day)	Energy Production (kWh)
January	0.687	0.687
February	0.768	0.768
March	0.8715	0.8715
April	0.966	0.966
May	1.0485	1.0485
June	1.068	1.068
July	1.0365	1.0365
August	0.9855	0.9855
September	0.8505	0.8505
October	0.8355	0.8355
November	0.756	0.756
December	0.6645	0.6645

Source: Authors, (2026).

VII.1.3 Cleanness Index

Clearer skies that are suitable for solar energy generation are indicated by values closer to 1 on the Cleanness Index, which measures sky purity.

• Highest Cleanness Index: With a Cleanness Index of 0.815, December has the clearest skies and the best solar-generating conditions.

• Lowest Cleanness Index: June has the lowest Cleanness Index (0.623) despite high solar radiation, indicating that air quality and cloud cover may have an impact on energy output.

VII.1.4 Seasonal Trends

• Spring to Summer: Solar radiation rises from January to June, reaching its maximum in June. In most locations, this cyclical pattern where summertime solar energy rises as the sun's angle becomes more direct—is common.
• Fall to Winter: Solar radiation levels decrease from June to December, with the lowest levels occurring in December. This decrease is due to reduced daylight hours and wintertime solar intensity.
• Cleanness Index Trends: During the summer, when there is usually more cloud cover, the Cleanness Index exhibits a similar seasonal tendency, peaking in December before declining.

VII.1.5 Energy Calculation

the annual energy production is calculated from a solar panel using the average daily solar radiation values for each month, assuming a panel area of 1 m² and an efficiency of 15%. The formula for calculating energy is as follows:
Esolar = Area × Radiation × Efficiency

Where:

• Area = 1 m² (for simplicity),
• Radiation = Monthly solar radiation values in kWh/m²/day,
• Efficiency = 0.15 (15%).

7.1.6. Solar Energy Calculation Results

Based on the provided monthly solar radiation values and assuming a 1 m² panel area with 15% efficiency, the monthly and total energy production is calculated as follows:
Monthly Energy Production (in kWh):
• January: 0.687 kWh
• February: 0.768 kWh
• March: 0.8715 kWh
• April: 0.966 kWh
• May: 1.0485 kWh
• June: 1.068 kWh
• July: 1.0365 kWh
• August: 0.9855 kWh
• September: 0.8505 kWh
• October: 0.8355 kWh
• November: 0.756 kWh
• December: 0.6645 kWh
Total Annual Energy Production: The total annual energy production from the solar panel is approximately 10.54 kWh.

VII.2 ANALYSIS OF MONTHLY AVERAGE WIND SPEED AND WIND ENERGY PRODUCTION

The Monthly Average Wind Speed data is analyzed in this study to assess the energy potential of a wind turbine in the area. The energy generated by a 3-meter-diameter wind turbine running at 35% efficiency, seasonal patterns, and wind speed breakdown are all included in the analysis. Figure (6) shows the data on wind speed and its seasonal fluctuations.

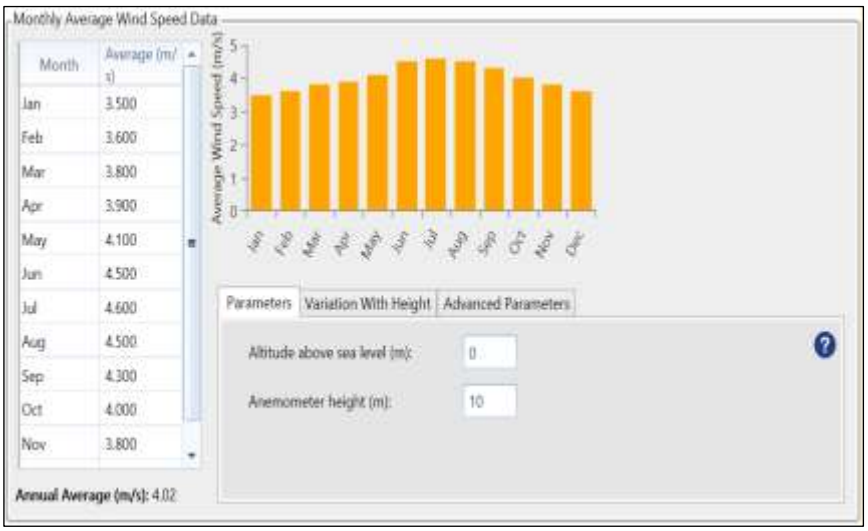


Figure 6: Analysis of Monthly Average Wind Speed Data.
Source: Authors, (2026).

VII.2.1 Annual Average Wind Speed

The annual average wind speed is 4.02 m/s, which indicates a moderate wind resource throughout the year.

VII.2.2 Monthly Breakdown of Wind Speed

As shown in Table 2, the monthly breakdown of wind speed (in m/s) and the corresponding energy production (in kWh) is presented. This data highlights the variation in wind speed and energy output throughout the year, indicating the periods of higher wind potential that can optimize energy generation.

Table 4: Monthly Breakdown of Wind Speed and Energy Production.

Month	Wind Speed (m/s)	Energy Production (kWh)
January	3.60	0.0707
February	3.60	0.0831
March	4.00	0.0899
April	4.10	0.1044
May	4.30	0.1381
June	4.40	0.1475
July	4.60	0.1381
August	4.30	0.1205
September	4.00	0.0970
October	3.80	0.0831
November	3.60	0.0707
December	3.60	0.0707

Source: Authors, (2026).

Seasonal Trends

• Spring and Summer (March to August): This season appears to be the most effective for generating wind energy because of the normally greater wind speeds, which peak in July at 4.60 m/s.
• Fall and Winter (September to February): During these colder months, wind speeds drop somewhat from 4.30 m/s to 3.60 m/s.
• Wind Energy Potential: The area has a moderate wind resource, with the greatest potential for wind energy generation occurring during the summer. As a result, wind turbines can operate effectively during these months.

Wind Energy Calculation

To calculate the energy production from a wind turbine, we use the following formula:
$E_{wind}=0.5 \times \rho \times A \times V^3 \times \eta$

Where:

ρ= Air density (1.225 kg/m³)
A= Swept area of the turbine blades ($A=\pi \times (D/2)^2$)
V = Wind speed (m/s)
η = Efficiency (35%)

The monthly energy production is calculated using the wind speed values for each month.

Wind Energy Production Calculation Results. Based on the monthly average wind speeds and the 3-meter diameter wind turbine with 35% efficiency, the monthly and total energy production is calculated as follows:

Monthly Energy Production (in kWh):

• January: 0.0707 kWh
• February: 0.0831 kWh
• March: 0.0899 kWh
• April: 0.1044 kWh
• May: 0.1381 kWh
• June: 0.1475 kWh
• July: 0.1381 kWh
• August: 0.1205 kWh
• September: 0.0970 kWh
• October: 0.0831 kWh

• November: 0.0707 kWh
• December: 0.0707 kWh

Total Annual Energy Production: The total annual energy production from the wind turbine is approximately 1.21 kWh.

VIII. CONCLUSION

Future power generation can be dependable with a hybrid solar-wind system. The Basra simulation confirmed solar dominance with supporting wind input, so validating their complementarity under local conditions. Global experiences, especially in China, Germany, and the US, demonstrate that hybrid systems frequently outperform single-source setups in terms of cost-effectiveness and reliability. Further developments in storage technology, intelligent control, and supportive legal frameworks are anticipated to make these systems essential in accelerating the transition to low-carbon and sustainable energy.

IX. FUTURE WORKS

Based on your research on solar-wind hybrid systems, here are some potential future works that could further expand and refine the field:

1. Optimization of Hybrid System Configurations: Future research could focus on developing advanced optimization models to determine the optimal size, configuration, and placement of solar and wind components within hybrid systems, considering local climate, weather patterns, and energy demands.
2. Advanced Energy Storage Solutions: Investigating the integration of next-generation energy storage technologies (e.g., solid-state batteries, flow batteries, compressed air energy storage) to improve the efficiency and reliability of hybrid systems, particularly in off-grid applications.
3. Smart Grid Integration and AI: Further exploration of how artificial intelligence (AI) and machine learning can be integrated into hybrid systems for real-time optimization and predictive maintenance, allowing for better energy management and grid stability.
4. Impact of Climate Change on Performance: Assessing how the growing effects of climate change (e.g., temperature fluctuations, shifting wind patterns, and changing sunlight availability) might influence the long-term performance and reliability of solar-wind hybrid systems.
5. Hybrid Systems in Urban Areas: Exploring the application of solar-wind hybrid systems in urban areas where space constraints are a challenge. Research could focus on integrating these systems with urban infrastructure (e.g., rooftops, vertical farms) for improved land use.
6. Sustainability of Manufacturing and Disposal: Researching the sustainability of materials used in hybrid energy systems, focusing on recycling and the reduction of environmental impacts from manufacturing and end-of-life disposal, particularly for wind turbine blades and solar panels.

X. REFERENCES

[1] Attanayake, K., Wickramage, I., Samarasinghe, U., Ranmini, Y., Ehalapitiya, S., Jayathilaka, R., et al., "Renewable energy as a solution to climate change: Insights from a comprehensive study across nations," PLOS ONE, vol. 19, no. 6, p. e0299807, Jun. 2024.

[2] Chilukuri, Y., Usman, A., Rajpurohit, B., "Assessing the Impact of Renewable Energy Sources to Achieve Net Zero Emissions," 2023 IEEE Conference on Technologies for Sustainability (SusTech), Portland, OR, USA, pp. 231-236, Apr. 2023.

[3] Q. Hassan, "A review of hybrid renewable energy systems: Solar and wind," Renewable Energy, vol. 182, pp. 1-13, 2023.

[4] International Energy Agency (IEA), "Renewables 2024 Electricity,"IEA,Paris,2024. https://www.iea.org/reports/renewables-2024/electricity

[5] S. Poddar, "Changes in Solar Resource Intermittency and Reliability,"ScienceDirect,2023.https://www.sciencedirect.com/science/article/pii/S0038092X2306734

[6] R. McKenna, "System Impacts of Wind Energy Developments," ScienceDirect, 2025. https://www.sciencedirect.com/science/article/pii/S2542435124005130

[7] M. J. Talukder, M. H. Hosen, and M. R. Khan, "Hybrid Renewable Energy Systems: Integrating Solar, Wind, and Biomass for Enhanced Sustainability and Performance," SSRN, 2022.

[8] A. Khan, M. Bressel, A. Davigny, D. Abbes, and B. Ould Bouamama, "Comprehensive Review of Hybrid Energy Systems: Challenges, Applications, and Optimization Strategies," Energies, vol. 18, no. 10, p. 2612, May 2025

[9] M. Savio, "Design of a Solar-Wind Hybrid Renewable Energy System," Engineering Reports, vol. 7, no. 5, p. e13101, 2025.

[10]A. H. Schleifer, D. Harrison-Atlas, W. J. Cole, and C. A. Murphy, "Hybrid renewable energy systems: The value of storage as a function of PV-wind variability," National Renewable Energy Laboratory, 2023.

[11] C. N. Nwagu, "Integrating solar and wind energy into the electricity grid for enhanced reliability," Renewable and Sustainable Energy Reviews, vol. 168, p. 112748, 2024.

[12]Yasser Fathi Nassar a, Hala Jarallah El-Khozondar, "The role of hybrid renewable energy systems in covering power shortages in public electricity grid: An economic, environmental and technical optimization analysis," Journal of Energy Storage Vol.108, 1 February 2025.

[13]Q. Hassan, S. Algburi, A. Z. Sameen, H. M. Salman, and M. Jaszczur, "A review of hybrid renewable energy systems: Solar and wind-powered solutions: Challenges, opportunities, and policy implications," Results in Engineering, vol. 20, p. 101621, Nov. 2023.

[14] J. Jurasz, F. A. Canales, A. Kies, M. Guezgouz, and A. Beluco, "A review on the complementarity of renewable energy sources: Concept, metrics, application and future research directions," Solar Energy, vol. 195, pp. 703-724, 2020.

[15] C. Murphy, D. Harrison-Atlas, N. Grue, T. Mosier, J. Gallego Calderon, and S. Elliott, "Complementarity of renewable energy-based hybrid systems," National Renewable Energy Laboratory, 2023.

[16] T. Jeyaraj, A. Ponnusamy, and D. Selvaraj, "Hybrid renewable energy systems stability analysis through future advancement technique: A review," Applied Energy, vol. 383, p. 125355, 2025, doi: 10.1016/j.apenergy.2025.125355.

[17] National Renewable Energy Laboratory (NREL), "Wind and Solar Hybrid Power Plants for Energy Resilience," NREL.

[18] A. A. Mas'ud, A. Yunusa-Kaltungo, N. A. Rufa'I, and N. Yusuf, "Assessing the viability of hybrid renewable energy systems in Nigeria," Engineering Reports, 2024.

[19] Kanaga Bharathi N, M. Abirami et al., "Techno-economic optimization of hybrid renewable systems for sustainable energy solutions," Nature Communications, 2025.

[20] Saini, K., Saini, M., Kumar, A. et al. Performance analysis and optimization in renewable energy systems: a bibliometric review. Discov Appl Sci 7, 178 (2025).

[21] B. F. Silinto, C. van der Laag Yamu, C. Zuidema, André P.C. Faaij, "Hybrid renewable energy systems for rural electrification in developing countries: A review on energy system models and spatial explicit modelling tools, Renewable
and Sustainable Energy Reviews," Volume 207, 2025, 114916, ISSN 1364-0321.

[22] S. Algburi, E. Rendal, Z. Kadum Jaber, H. F. Fakhruldeen, A. Amjad, I.B. Sapaev, T. T. Farhan, M. Barakat, "Evaluating the Viability and Potential of Hybrid Solar-Wind Renewable Energy Systems in Relation to Geographical and Environmental Factors," ScienceDirect, 2024.

[23] J. B. Fulzele and M. B. Daigavane, "Design and optimization of hybrid PV-wind renewable energy system," Materials Today: Proceedings, vol. 5, no. 1, Part 1, pp. 810-818, 2018. doi: 10.1016/j.matpr.2017.11.151.

[24] Wang G, Zhang Y, Tang W, Liao Z, Wang T, Zhang S and Zhao X, "Feasibility analysis of hybrid energy generation systems for service areas along desert highways," Frontiers in Energy Research, Vol. 12 2024.

[25] Solar and wind power generation, 2024

[26] N, K.B., Abirami, M., Vighneshwari, D. et al. "Techno-economic optimization of hybrid renewable systems for sustainable energy solutions," Sci Rep 15, 21114 (2025).

[27] H. Al-Rawashdeh, O. A. Al-Khashman, J. T. Al Bdour, M. R. Gomaa, H. Rezk, A. Marashli, L. M. Arrfou, and M. Louzazni, "Performance analysis of a hybrid renewable-energy system for green buildings to improve efficiency and reduce GHG emissions with multiple scenarios," Sustainability, vol. 15, no. 9, p. 7529, 2023. doi: 10.3390/su15097529.

[28] A. Rama Krishna, A. Vijay Kumar, A. Gopala Krushna, J. Shanmugapriyan, and C. Keertana, "The study of solar and wind power systems under different weather conditions," E3S Web of Conferences, vol. 547, p. 03009, 2024.

[29] Q. Hassan, S. Algburi, A. Z. Sameen, H. M. Salman, and M. Jaszczur, "A review of hybrid renewable energy systems: Solar and wind-powered solutions: Challenges, opportunities, and policy implications," Results in Engineering, vol. 20, p. 101621, 2023.

[30] A. Alkhafa, M. G. Kadhim, F. A. Alhaddad, and A. Saad, "Performance analysis of hybrid renewable energy systems under variable operating conditions," Solar Compass, vol. 15, p. 100134, 2025.

[31] S. R. Golroudbary, I. Makarava, A. Kraslawski, and E. Repo, "Global environmental cost of using rare earth elements in green energy technologies," Science of The Total Environment, vol. 832, p. 155022, Aug. 2022, doi:
10.1016/j.scitotenv.2022.155022.

[32] Bamisile, O., Acen, C., Cai, D., Huang, Q., & Staffell, I. (2024). The environmental factors affecting solar photovoltaic output. Renewable and Sustainable Energy Reviews,208,115073.https://doi.org/10.1016/j.rser.2024.115073

[33] Bonkaney, A., Madougou, S., & Adamou, R. (2017). Impacts of cloud cover and dust on the performance of photovoltaic module in Niamey. Journal of Renewable Energy, 2017, 1–8. https://doi.org/10.1155/2017/9107502

[34] Zhang, B., Cheng, S., Lu, F., Zheng, Y., & Chamorro, L. P. (2020). Impact of topographic steps in the wake and power of a wind turbine: Part A—Statistics. Energies, 13(23), 6411. https://doi.org/10.3390/en13236411

[35] Torres, J. F., & Petrakopoulou, F. (2022). A closer look at the environmental impact of solar and wind energy. Global Challenges, 6(8). https://doi.org/10.1002/gch2.202200016

[36] Moustakas, A., Georgiakakis, P., Kret, E., & Kapsalis,E. (2023, May 30). Wind turbine power and land cover effects on cumulativ ebat deaths. arXiv. org. https://arxiv.org/abs/2305.19010

[37] D. Roy, K. V. Shivaprasad, Y. Wang, and A. P. Roskilly, "Techno-economic feasibility assessment for a standalone hybrid energy system integrating renewable energy sources," Cleaner Eng. Technol., vol. 28, p. 101045, 2025, doi:
10.1016/j.clet.2025.101045.

[38] Jadhav, N. B., Gajare, O., Zele, S., Gogate, N., & Joshi, A. (2024). Current status and challenges in silver recovery from End-of-Life crystalline silicon solar photovoltaic panels. Solar Energy, 283, 113027. https://doi.org/10.1016/j.solener.2024.113027

[39] Wind turbine blades: Options at end of life - Kleinman Center for Energy Policy. (2024, September 18). Kleinman Center for Energy Policy.

[40] J. Guan, "Landscape visual impact evaluation for onshore wind farm: A case study," ISPRS Int. J. Geo-Inf., vol. 11, no. 12, p. 594, 2022. https://doi.org/10.3390/ijgi11120594

[41] J. Abromas, P. Grecevičius, and N. Piekienė, VISUAL IMPACT ASSESSMENT OF WIND TURBINES ON LANDSCAPE IN ŠILALĖ REGION. 2015. doi: 10.15544/rd.2015.005.

[42] Y. Xiao, C. Zou, M. Dong, H. Chi, Y. Yan, and S. Jiang, "Feasibility study: Economic and technical analysis of optimal configuration and operation of a hybrid CSP/PV/wind power cogeneration system with energy storage," Renewable Energy, vol. 225, p. 120273, 2024.

[43] M. M. Ibrahim, "Energy management strategies of hybrid renewable energy systems: A review," Wind Engineering, vol. 48, no. 1, pp. 133–161, Oct. 2023, doi: 10.1177/0309524x231200010.

[44] C. R. Kumar. J and M. A. Majid, "Renewable energy for sustainable development in India: current status, future prospects, challenges, employment, and investment opportunities," Energy Sustainability and Society, vol. 10, no. 1, Jan. 2020, doi: 10.1186/s13705-019-0232-1.

[45] P. R. Brown, T. Williams, M. L. Brown, and C. Murphy, "System-cost-minimizing deployment of PV-wind hybrids in low-carbon U.S. power systems," Applied Energy, vol. 365, p. 123151, Apr. 2024, doi: 10.1016/j.apenergy.2024.123151.

[46] R. A. A. Hasibi and A. Haris, "An analysis of the implementation of a hybrid renewable-energy system in a building by considering the reduction in electricity price subsidies and the reliability of the grid," Clean Energy, vol. 7, no. 5, pp. 1125–1135, Oct. 2023, doi: 10.1093/ce/zkad053.

[47] S. Pathak, "Hybrid renewable energy: paving the way for a greener tomorrow," Hero Future Energies, Apr. 16, 2025. https://www.herofutureenergies.com/blog/hybrid-renewable-energy/?utm_source=chatgpt.com

[48] Qadir, S. A., Al-Motairi, H., Tahir, F., & Al-Fagih, L. (2021). Incentives and strategies for financing the renewable energy transition: A review. Energy Reports, 7, 3590–3606. https://doi.org/10.1016/j.egyr.2021.06.041

[49] https://www.nrel.gov/state-local-tribal/energy-compensation-mechanisms

[50] A. Altassan, "Sustainable integration of solar energy, behavior change, and recycling practices in educational institutions: A holistic framework for environmental conservation and quality education," Sustainability, vol. 15, p. 15157, 2023.

[51] https://www.cleanenergywire.org/factsheets/qa-how-will-germany-support-expansion-renewables-future

[52] Qadir, S. A., Al-Motairi, H., Tahir, F., & Al-Fagih, L. (2021). Incentives and strategies for financing the renewable energy transition: A review. Energy Reports, 7, 3590–3606. https://doi.org/10.1016/j.egyr.2021.06.041

[53] Olbrich, S., Bauknecht, D., & Späth, P. (2024). Policy mixes for net-zero energy transitions: Insights from energy sector integration in Germany. Energy Research & Social Science, 118, 103822. https://doi.org/10.1016/j.erss.2024.103822

[54] Borenstein, S., & Davis, L. W. (2016). The distributional effects of US clean energy tax credits. Tax Policy and the Economy, 30(1), 191–234. https://doi.org/10.1086/685597

[55] Shen, J., & Luo, C. (2014). Overall review of renewable energy subsidy policies in China Contradictions of intentions and effects. Renewable and Sustainable Energy Reviews,41,1478–1488. https://doi.org/10.1016/j.rser.2014.09.007

[56] Das, A., Jani, H. K., Nagababu, G., & Kachhwaha, S. S. (2020). A comprehensive review of wind–solar hybrid energy policies in India: Barriers and Recommendations. Renewable EnergyFocus,35,108–121. https://doi.org/10.1016/j.ref.2020.09.004

[57] De Andrade Santos, J. a. F., De Jong, P., Da Costa, C. A., & Torres, E. A. (2020). Combining wind and solar energy sources: Potential for hybrid power generation in Brazil. Utilities Policy, 67, 101084. https://doi.org/10.1016/j.jup.2020.101084

[58] Numbi, B., & Malinga, S. (2016). Optimal energy cost and economic analysis of a residential grid-interactive solar PV system- case of eThekwini municipality in South Africa. Applied Energy, 186, 28–45. https://doi.org/10.1016/j.apenergy.2016.10.048

[59] L. E. Natividad and P. Benalcazar, "Hybrid Renewable Energy Systems for Sustainable Rural Development: Perspectives and Challenges in Energy Systems Modeling," Energies, vol. 16, no. 3, p. 1328, 2023.

[60] Alex-Oke, T., Bamisile, O., Cai, D., Adun, H., Ukwuoma, C. C., Tenebe, S. A., & Huang, Q. (2025). Renewable energy market in Africa: Opportunities, progress, challenges, and future prospects. Energy Strategy Reviews, 59, 101700. https://doi.org/10.1016/j.esr.2025.101700

[61] Zebra, E. I. C., Van Der Windt, H. J., Nhumaio, G., & Faaij, A. P. (2021). A review of hybrid renewable energy systems in mini-grids for off-grid electrification in developing countries. Renewable and Sustainable Energy Reviews, 144, 111036. https://doi.org/10.1016/j.rser.2021.111036

[62] Numbi, B., & Malinga, S. (2016). Optimal energy cost and economic analysis of a residential grid-interactive solar PV system- case of eThekwini municipality in South Africa. Applied Energy, 186, 28–45. https://doi.org/10.1016/j.apenergy.2016.10.048

[63] Slimene, M. B., & Khlifi, M. A. (2025). A hybrid renewable energy system with advanced control strategies for improved grid stability and power quality. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-06091-w

[64] Kumari, N., Tran, B., Sharma, A., & Alahakoon, D. (2025). A Comprehensive Review on stability analysis of hybrid Energy System. Sensors, 25(10), 2974. https://doi.org/10.3390/s25102974

[65] Savio, F. M., Joshua, S. V., Usha, K., Faheem, M., Kannadasan, R., & Khan, A. A. (2025c). Design of a Solar-Wind hybrid renewable energy system for power quality enhancement: a case study of 2.5 MW real time domestic grid.
Engineering Reports,7(1). https://doi.org/10.1002/eng2.13101

[66] Jose, S. A., Gallant, A., Gomez, P. L., Jaggers, Z., Johansson, E., LaPierre, Z., & Menezes, P. L. (2025). Solid-State Lithium Batteries: Advances, challenges, and future perspectives. Batteries, 11(3), 90. https://doi.org/10.3390/batteries11030090

[67] Sarra, Z., Bouziane, M., Bouddou, R., Benbouhenni, H., Mekhilef, S., & Elbarbary, Z. (2024). Intelligent control of hybrid energy storage system using NARX-RBF neural network techniques for microgrid energy management. Energy Reports, 12, 5445–5461. https://doi.org/10.1016/j.egyr.2024.11.023

[68] Mamodiya, U., Kishor, I., Garine, R., Ganguly, P., & Naik, N. (2025). Artificial intelligence based hybrid solar energy systems with smart materials and adaptive photovoltaics for sustainable power generation. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-01788-4

[69] Falope, T., Lao, L., Hanak, D., & Huo, D. (2024). Hybrid energy system integration and management for solar energy: A review. Energy Conversion and Management X, 21, 100527. https://doi.org/10.1016/j.ecmx.2024.100527

[70] Dranka, G. G., Ferreira, P., & Vaz, A. I. F. (2021). Integrating supply and demand-side management in renewable-based energy systems. Energy, 232, 120978. https://doi.org/10.1016/j.energy.2021.120978

[71] Sayed, E. T., Wilberforce, T., Elsaid, K., Rabaia, M. K. H., Abdelkareem, M. A., Chae, K., & Olabi, A. (2020). A critical review on environmental impacts of renewable energy systems and mitigation strategies: Wind, hydro, biomass
and geothermal. The Science of the Total Environment, 766, 144505. https://doi.org/10.1016/j.scitotenv.2020.144505

[72] Muller, D. C., Selvanathan, S. P., Cuce, E., & Kumarasamy, S. (2023).

[73] Hybrid solar, wind, and energy storage system for a sustainable campus: A simulation study. Science and Technology for Energy Transition, 78, 13.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 147-155. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2830

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

FAULT PROTECTION ANALYSIS OF A DC MICROGRID USING VSC BASED MODULAR MULTILEVEL CONVERTER (MMC)

Kanna Sravanthi*[1] and Ayya Chaithanya[2]

[1, 2]Department of Electrical and Engineering, SV University, Tirupati, India.

[1]https://orcid.org/0009-0000-5910-7169, [2]https://orcid.org/0009-0004-7501-210X

Email: *sravanthikanna7@gmail.com, chaithu.maruthi@gmail.com

ARTICLE INFO

Article History
Received: October 17, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
AC Microgrid,
DC Microgrid,
PV System,
Voltage Source Converter,
Modular Multilevel Converter,
Differential Relay

ABSTRACT

DC microgrid development is increasing mainly due to its simplicity of operation and greater focus on renewable energy sources like solar, wind, etc. It comprises less losses than AC microgrid and can easily connect and disconnect distributed generation (DG). Conventional techniques of DC microgrid protection is challenging. Using MATLAB Simulink, a DC microgrid with a PV system, battery, and AC main grid along with converters including boost, buck, and voltage source/modular multilevel converters is simulated. This study compares the performance of modular multilevel converters and voltage source converters, as well as the way they protect DC microgrid in the case of an AC main grid fault condition. Additionally, for improved fault clearing response in overcurrent situations, the differential relay is used on the AC main grid, nearer to the transformer.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Nowadays, grid integration and renewable energy sources are the primary areas of research study. Due to fossil fuels are becoming extinct and their emissions affect the environment, the usage of renewable energy sources has been growing. Decentralized generation (DG) is connected using both island/isolated and grid-connected techniques [1]. The grid-connected state has significant advantages over the island state. In grid-connected state, the grid provides power to the load is able to meet its demand in the event that there is any problem with the decentralized generation (DG). Depending on the type of system, converters including inverters, rectifiers, and filters are required for connecting decentralized generation (DG) to the grid. An extensive collection of decentralized generation systems is called a microgrid [2]. Compared to AC microgrids, DC microgrids are more effective, like the absence of reactive power and synchronization problems [3], [4]. DC microgrids have greater power transfer capabilities [5]. A DC microgrid's power flow is more robust due to multiple sources can connect to a single bus [6].

To integrate a AC main grid with the DC microgrid, a converter is required. These include multilevel converters, voltage source converters (VSC) and current source converters (CSC) [7]. Multilevel converters are utilized for large power applications, while voltage source converters operate efficiently for small and medium power applications. Current source converters (CSC's) need external AC capacitors for reactive power compensation and required forced or line commutation. VSCs are the most compatible and widely used topologies. But due to its complex construction and lack of adaptability when extending the voltage or power range, VSC is not suitable for high power applications. For high power applications modular multilevel converter is used. Capacitors and at the DC bus, additional passive filters are not necessary with MMCs, which is an advantage [8]. For DC Microgrid protection, the differential overcurrent relay is used near to the transformer for precise fault detection [9].

II. FAULT PROTECTION ANALYSIS

On a microgrid, the current direction changed whenever a fault occurred [10]. Circuit breakers receive the signal from relays and turn off the system that is out of fault. The DC Microgrid current direction before and after the fault develops as shown in the diagram below [11], [12].

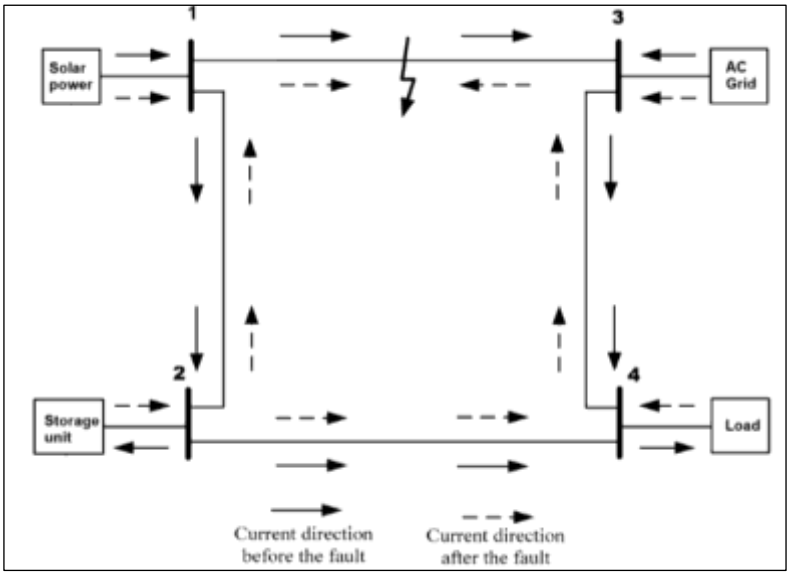


Figure 1: Before and after fault DC Microgrid current direction.
Source: Authors, (2026).

The VSC and MMC act as switch between the AC main and the DC microgrid, as shown below [9].

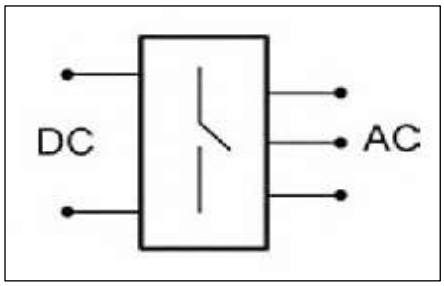


Figure 2: VSC/MMC acts as switch between DC and AC system.
Source: Authors, (2026).

The circuit design for the MMC and VSC are shown below. Capacitors are present in the Modular Multilevel Converter's (MMC) sub module block. In order to safeguard the DC Microgrid against faults, the MMC does not require any external large capacitors [13-15].

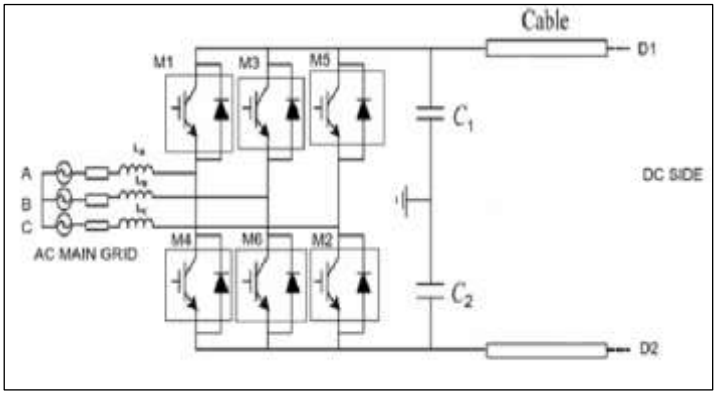


Figure 3: VSC circuit diagram.

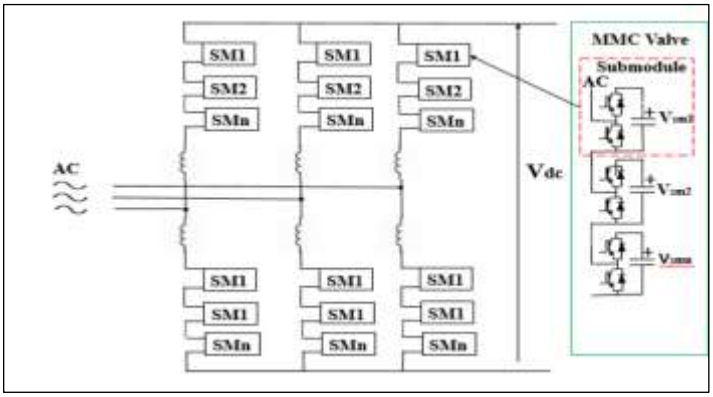


Figure 4: MMC circuit diagram.

Source: Authors, (2026).

The three cases below shows the fault that occurred on the DC side. The condition before the fault occurred was t_o. At this point, the cable current is zero and the capacitor reaches its full charge. At t_o, the fault occurred, the cable current is gradually increasing, and the capacitor's stored charge is gradually decreasing. When the capacitor's stored charge reaches zero at t_1, the grid gradually begins feeding. The charging of the capacitor starts at t_2 [16].

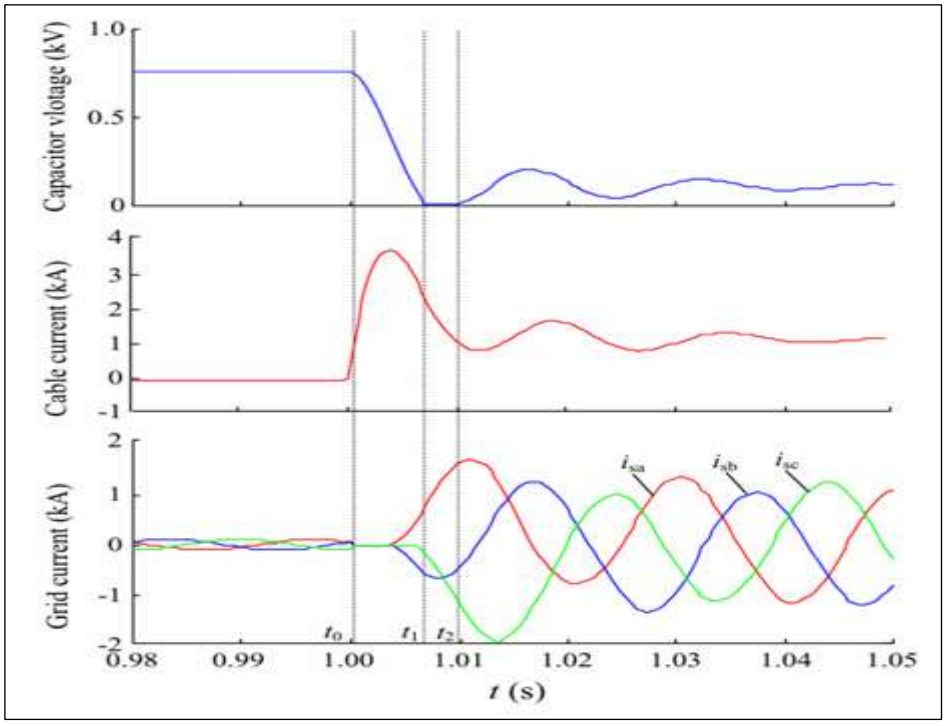


Figure 5: Fault characteristics.
Source: Authors, (2026).

III. SIMULATION

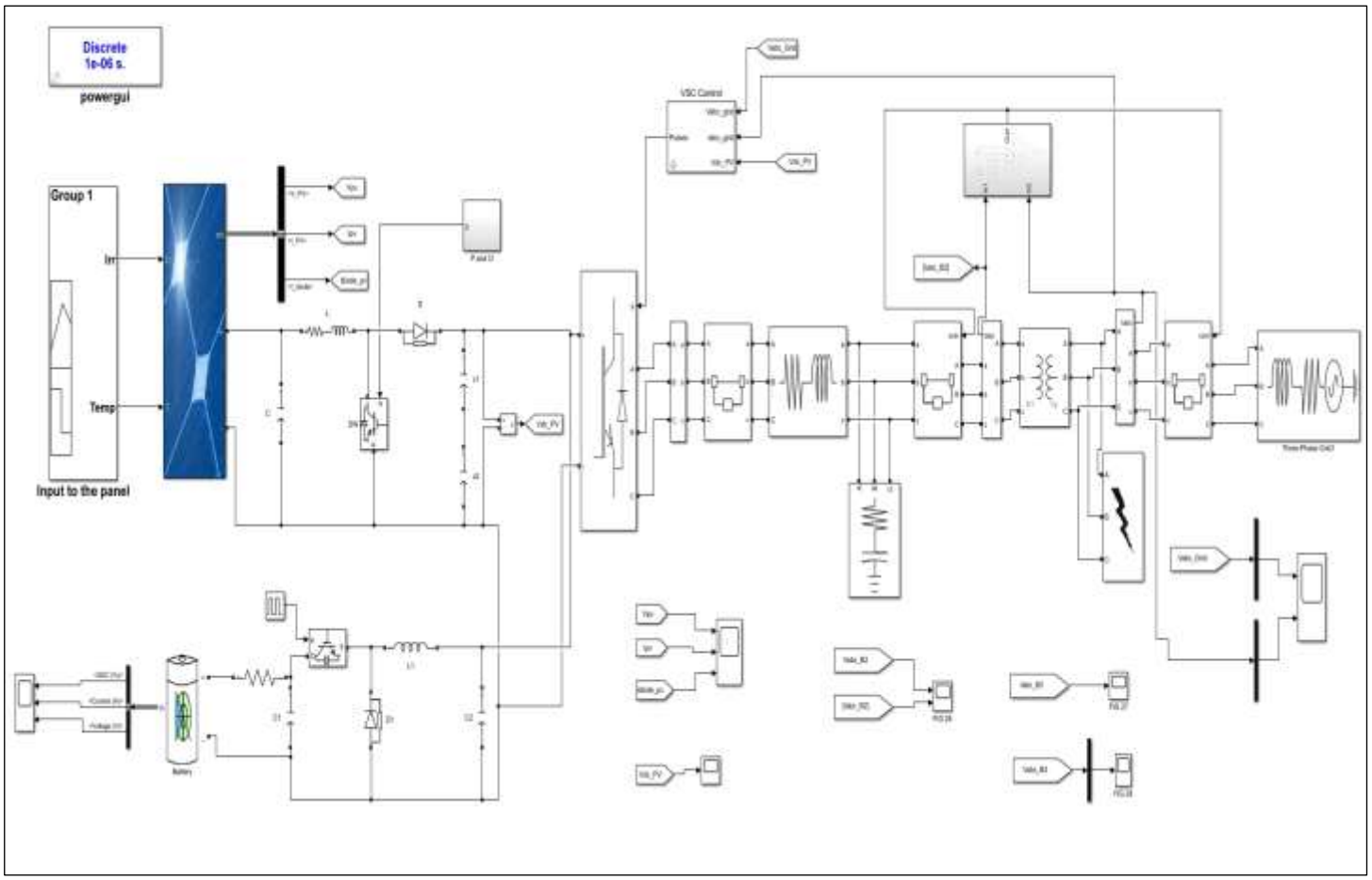


Figure 6: Simulink design of DC Microgrid and AC Main Grid with Voltage Source Converter (VSC).
Source: Authors, (2026).

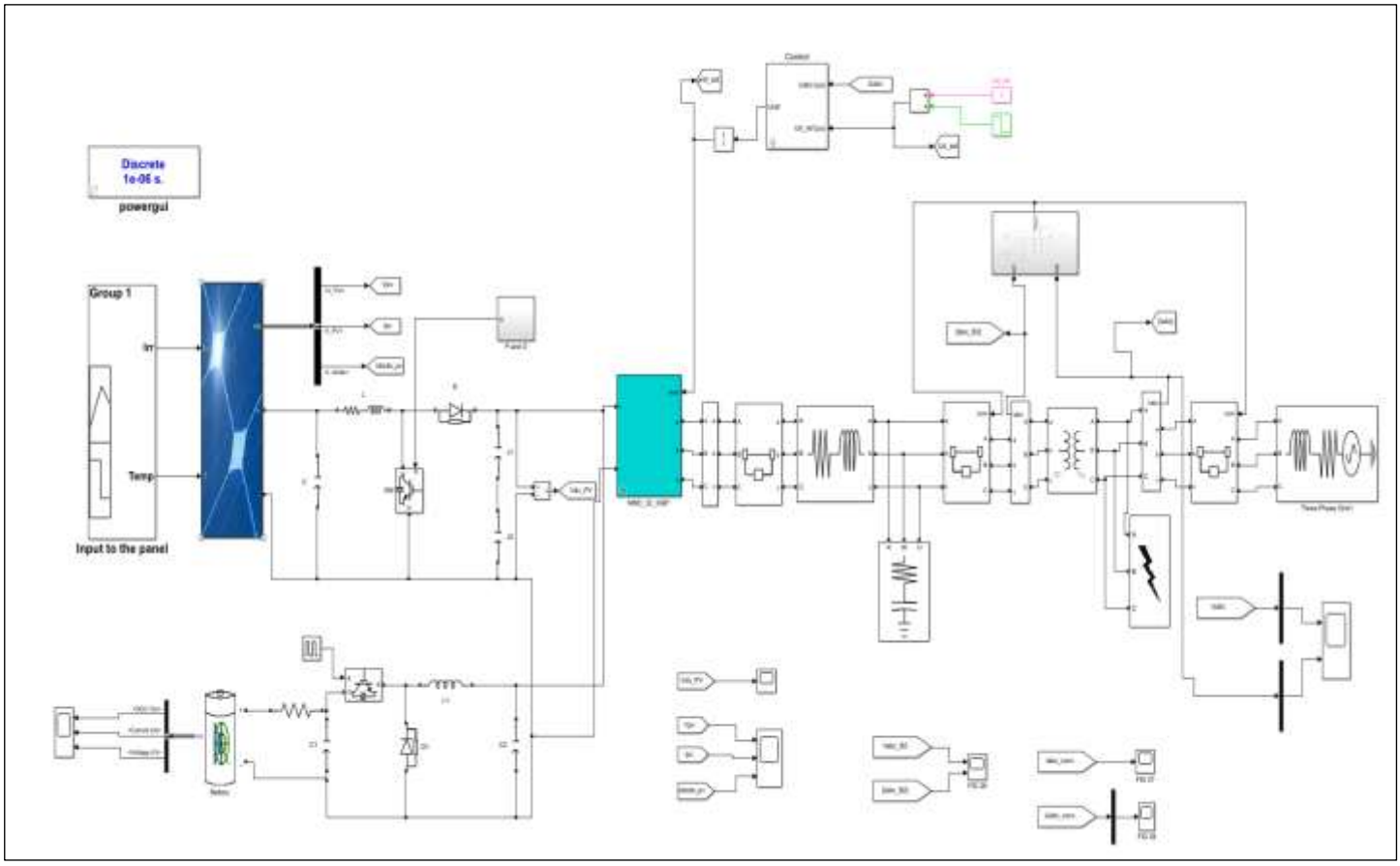


Figure 7: Simulink design of DC Microgrid and AC Main Grid with 32 IGBT Modular Multilevel Converter (MMC).
Source: Authors, (2026).

Table 1: AC Main Grid specifications.

S.No	Name of parameters	Value
1	Phase to phase voltage (V_{rms}) of three-phase Grid	25 KV
2	Nominal power of three-phase transformer	100 KVA
3	Active power of three phase series RLC load	10 KW
4	Nominal phase to phase voltage (V_{rms})	440 V

Source: Authors, (2026).

IV. RESULTS AND DISCUSSIONS

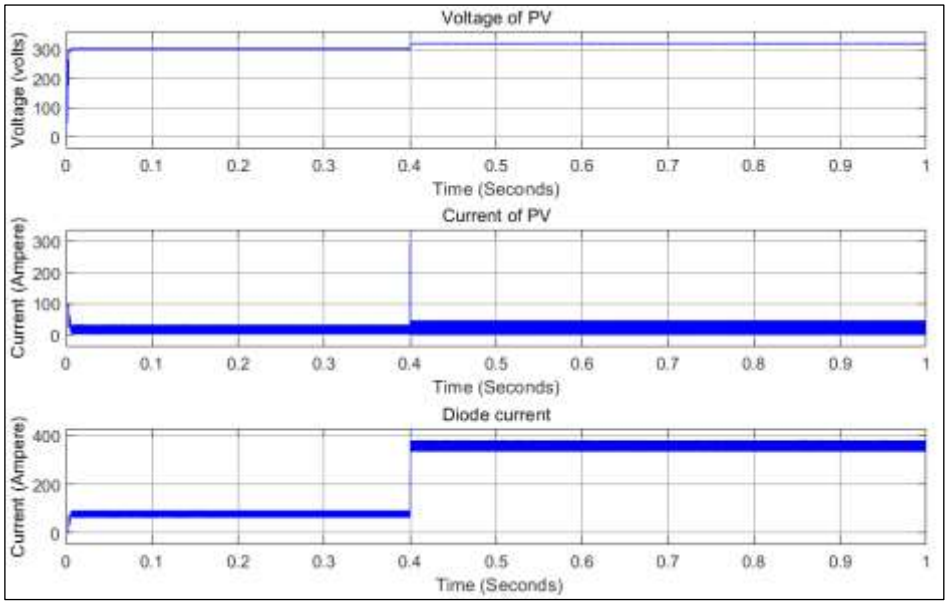


Figure 8: Voltage, current and diode current of PV system.
Source: Authors, (2026).

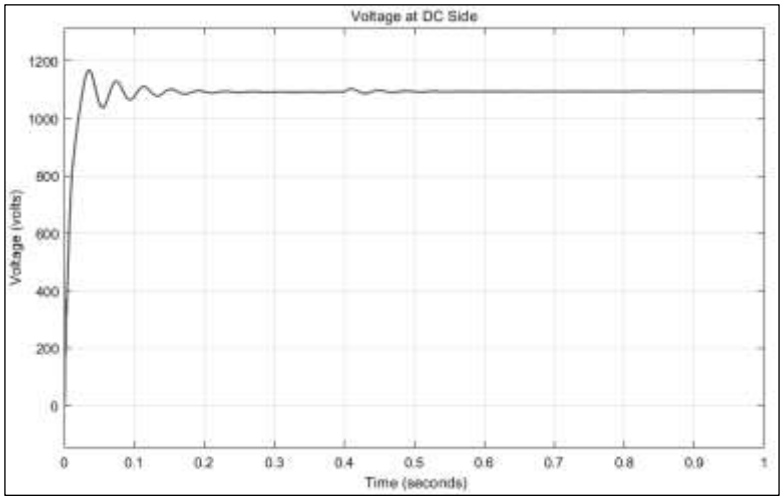


Figure 9: Output voltage of DC Microgrid.
Source: Authors, (2026).

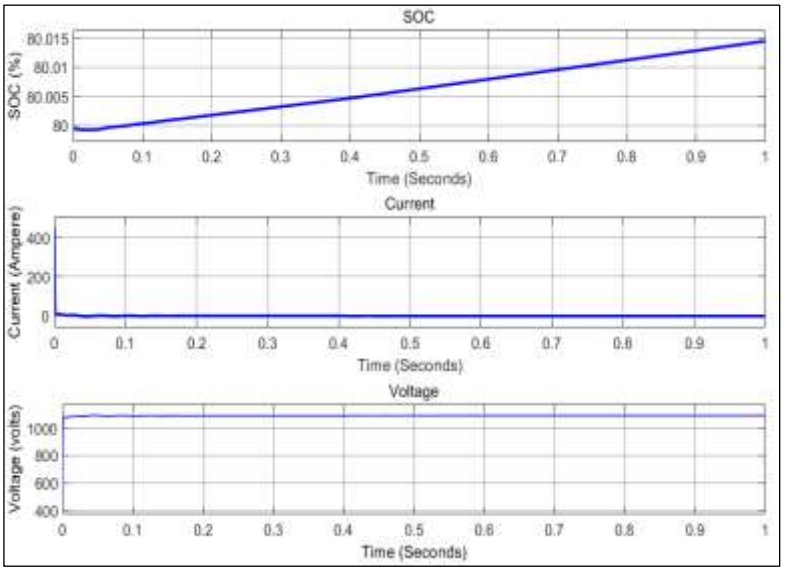


Figure 10: Battery parameters.
Source: Authors, (2026).

The figures mentioned above display the voltage, current, and diode current of the PV system. A DC microgrid's output voltage is 1.169KV. Lithium-ion battery parameters are displayed.

IV.1 FAULT CONDITION (LLLG) OF VSC

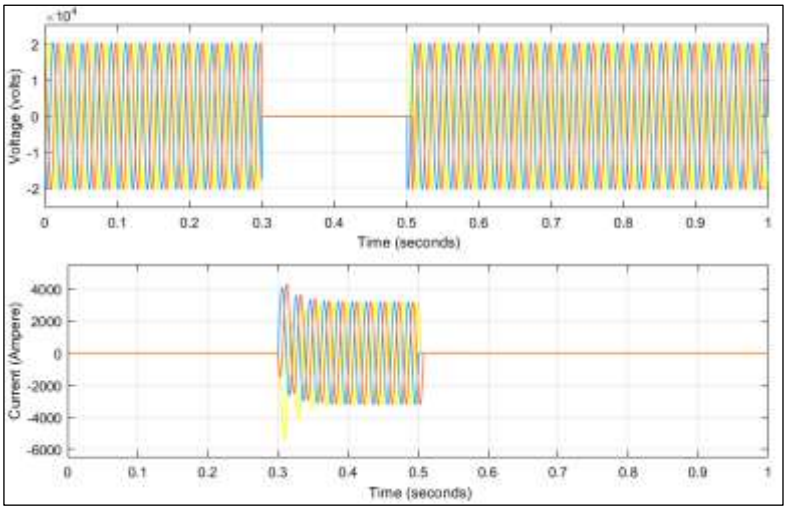


Figure 11: Voltage and current of AC Main Grid by using VSC.
Source: Authors, (2026).

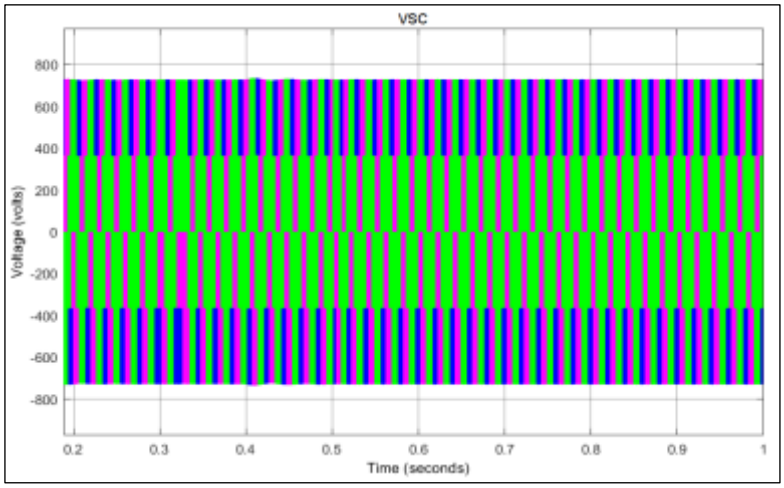


Figure 12: Voltage of VSC.
Source: Authors, (2026).

The LLLG fault is applied between 0.3 and 0.5 seconds on the AC Main Grid. Variations in AC Main Grid voltage and current are noted before to, during, and after fault clearance conditions. But the fault does not much interpret the voltage of Voltage Source Converter. In this way, the Voltage Source Converter protects the DC microgrid from faults.

IV.1.1 Fault Condition (LLLG) of MMC

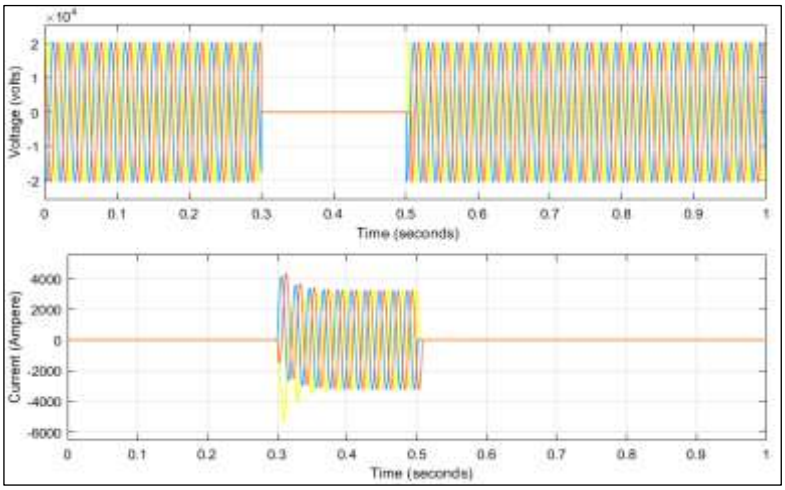


Figure 13: Voltage and current of AC Main Grid by using MMC.
Source: Authors, (2026).

The LLLG fault is applied between 0.3 and 0.5 seconds on the AC Main Grid. Variations of voltage and current in AC Main Grid are noted before, during, and after fault clearance conditions. But the fault does not much interpret the voltage of the modular multilevel converter.

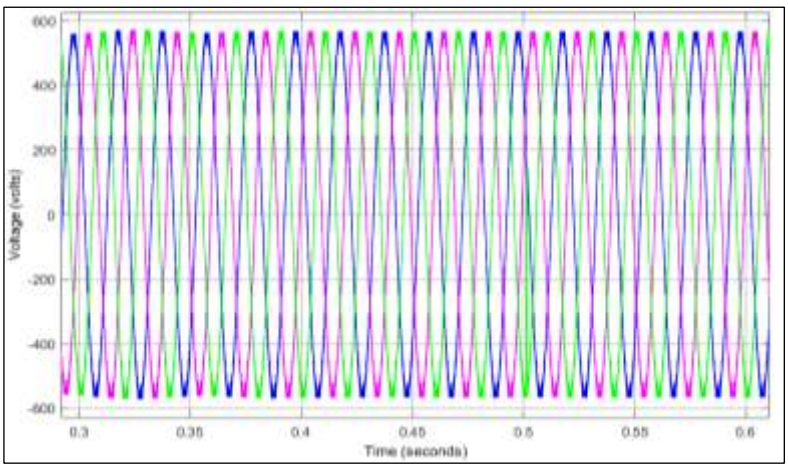


Figure 14: Voltage of MMC.
Source: Authors, (2026).

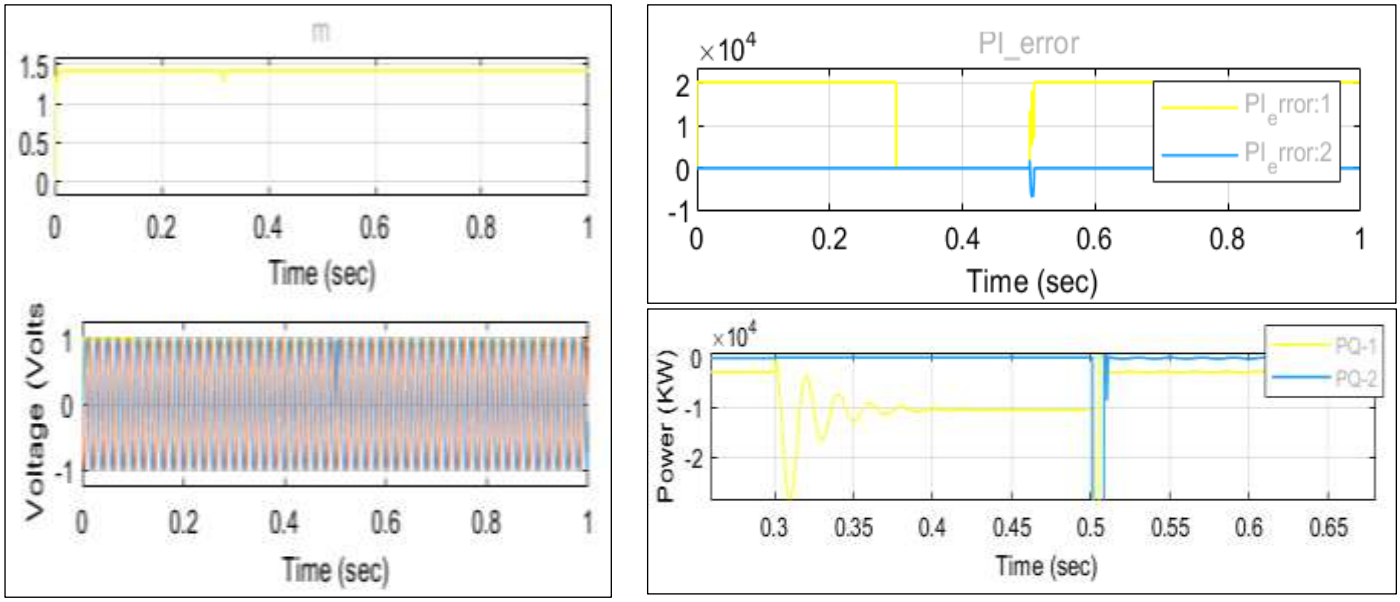


Figure 15: Waveforms of power and modulation index.
Source: Authors, (2026).

An inverter with six switches three for the positive arm and three for the negative arm is created using MMC circuit 32 IGBT. When the model is ready, it runs for one second, and on the AC Main Grid side, fault is imposed at time periods of 0.3 and 0.5 seconds. For voltage regulation, an MMC based on VSC is used. The triggering signal is produced by a triangle waveform generated by MMC with PWM [17], and the reference voltage is controlled by a circuit. The PWM input receives a reference voltage of 1pu that is generated in the controlling structure.

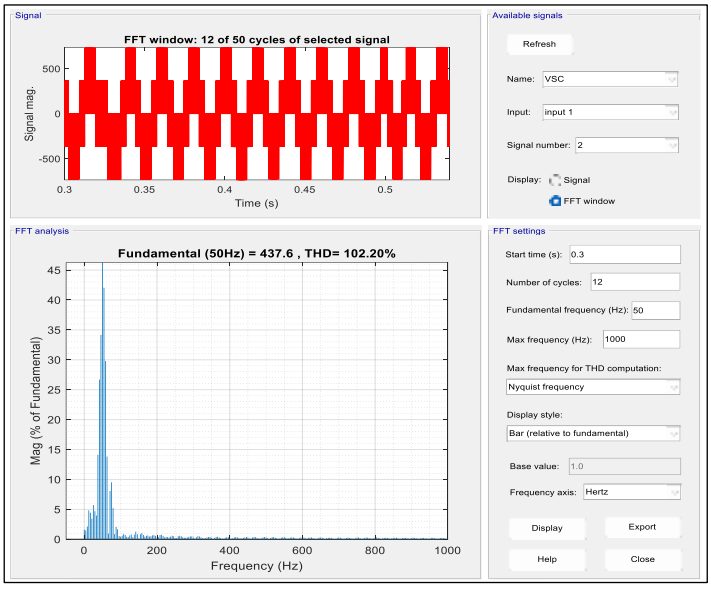


Figure 16: FFT Plot of VSC.
Source: Authors, (2026).

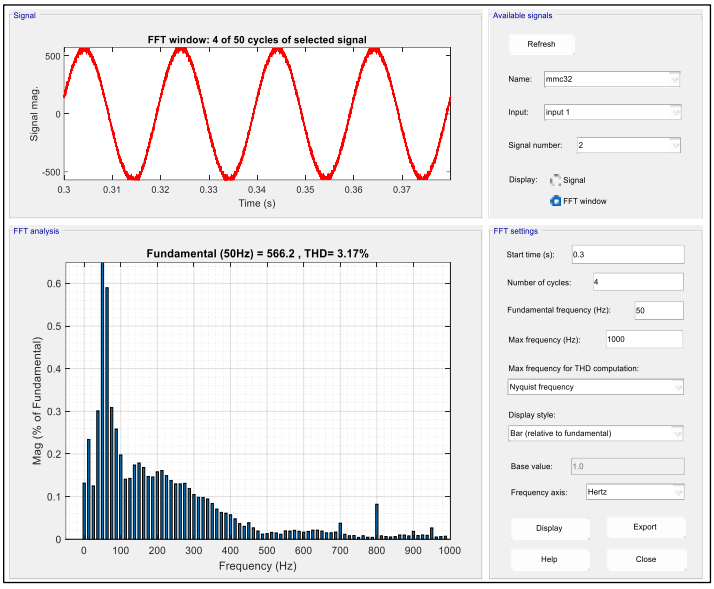


Figure 17: FFT Plot of MMC.
Source: Authors, (2026).

IV.1.2 Comparative Study

The Comparison of (THD)% of VSC and MMC is shown below.

Table 2: Comparison of (THD)% of VSC and MMC.

Parameters	VSC	MMC
Total Harmonic Distortion (THD)%	102.02	3.17
Fundamental Frequency (50Hz)	437.6	566.2

Source: Authors, (2026).

From the comparative study, the total harmonic distortion (THD) percentage of the voltage source converter (VSC) is 102.02%, whereas that of the modular multilevel converter (MMC) is 3.17%. The Modular Multilevel Converter (MMC) lowers harmonic frequencies. In terms of DC Microgrid protection, this indicates that the modular multilevel converter performs better than the voltage source converter.

V. CONCLUSIONS

MATLAB Simulink is used to analyze the DC microgrid's fault protection analysis using a VSC and a MMC. When compared to a voltage source converter, the THD analysis of a 32-level IGBT modular multilevel converter based on VSC exhibits good results. When the AC main grid has a fault, the 32-level IGBT modular multilevel converter based on VSC operates better at protecting the DC microgrid. The Power Electronics Building Block model is used due of the system's highly modular design and execution. Only a differential relay is used for fault protection and a PI controller is used. Neural network ANFIS converters, various relays and filters can be used further.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Kanna Sravanthi and Ayya Chaithanya.
Methodology: Kanna Sravanthi and Ayya Chaithanya.
Investigation: Kanna Sravanthi and Ayya Chaithanya.
Discussion of results: Kanna Sravanthi and Ayya Chaithanya.
Writing – Original Draft: Kanna Sravanthi.
Writing – Review and Editing: Kanna Sravanthi and Ayya Chaithanya.
Resources: Kanna Sravanthi.
Supervision: Ayya Chaithanya.
Approval of the final text: Kanna Sravanthi and Ayya Chaithanya.

VII. ACKNOWLEDGMENTS

The author is grateful to A. Chaithanya of SV University's Electrical & Electronics Engineering Department for her guidance in bringing this paper to completion. Also, the author likes to show thankfulness to all of the educational professionals in the electrical and electronics engineering Department in SV University who gave either direct or indirect support in completing this paper

VIII. REFERENCES

[1] L. Zhang, N. Tai, W. Huang, J. Liu and Y. Wang, "A review on the protection of DC microgrids," in Journal of Modern Power Systems and Clean Energy, vol. 6, no. 6, pp. 1113-1127, November 2018.

[2] Priya Ranjan Satpathy., "Advanced protection technologies for microgrids: Evolution, challenges, and future trends," in Energy Strategy Reviews Volume 58, March 2025, 101670.

[3] Akshaykumar N. Wankhade, Prof. Roshan R. Deotare, "Protection scheme for unipolar mmc-based multi-terminal HVDC grid," 2021 JETIR July 2021, Volume 8, Issue 7.

[4] T.Strasser, F. Andrén, J. Kathan, C. Cecati, C. Buccella, P. Siano, P. Leitão, G. Zhabelova, V. Vyatkin, P. Vrba, and V. Maríık, ''A review of architectures and concepts for intelligence in future electric energy systems,'' IEEE Trans. Ind. Electron., vol. 62, no. 4, pp. 2424–2438, Apr. 2015.

[5] Mohammed Ahmed Yagoub., "Modelling and Fault Protection Analysis of a DC Microgrid using Voltage Source Converter (VSC)," in 2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE) | 978-1-7281-9111-9/20/$31.00 ©2021 IEEE|.

[6] J. Y. Kim et al., "Cooperative Control Strategy of Energy Storage System and Microsources for Stabilizing the Microgrid during Islanded Operation," in IEEE Transactions on Power Electronics, vol. 25, no. 12, pp. 3037-3048, Dec. 2010.

[7] Immanuel N. Jiya, Ahmed Salem, "Integrated Multiport DC-DC and Multilevel Converters for Multiple Renewable Energy Source Integration," IEEE Access, Digital Object Identifier 10.1109/ACCESS.2023.3336567, volume 11, 2023.

[8] Rainer Marquardt, "Modular Multilevel converters State of the art and future progress," IEEE Power Electronics Magazine, 2018.

[9] Ming Yu, Yi Wang, Lirong Zhang, Ziguang Zhang. "DC short circuit fault analysis and protection of ring type DC microgrid", 2016 IEEE 8th International Power Electronics and Motion Control Conference (IPEMC-ECCE Asia), 2016.

[10] C. A. Apostolopoulos and G. N. Korres, "A Novel Algorithm for Locating Faults on Transposed/Untransposed Transmission Lines Without Utilizing Line Parameters," in IEEE Transactions on Power Delivery, vol. 25, no. 4, pp. 2328-2338, Oct. 2010.

[11] Sumanth, Akash Srivastava and Sangeeta Modi, "Fault Analysis and Protection of DC Microgrid," International Journal of Engineering Research & Technology (IJERT), volume 11, pp. 122-126, IETE-2020.

[12] EYAN, Abhisek UKIL, "Simulation and Analysis of Faults in High Voltage DC Power Transmission," Nanyang Technological University, Singapore, 2014.

[13] K. Sreedhar Kumar, J N Chandra Sekhar, "Limitation of Fault Currents by Positionind of Superconducting Fault Current Limiter in AC and DC Microgrid" IEEE Trans.Power Del.,vol. 4, no. 2, February. 2015. Maanickam KARTHIK.

[14] R. M. Cuzner and G. Venkataramanan, "The Status of DC Micro-Grid Protection," 2008 IEEE Industry Applications Society Annual Meeting, Edmonton, AB, 2008, pp. 1-8.

[15] Jassem Mohammed., "Performance Analysis of Overcurrent Protection Schemes In Power Systems," in 2024 Advances in Science and Engineering Technology International Conferences (ASET) | 979-8-3503-4413-4/24/$31.00 ©2024 IEEE |.

[16] N. Bayati, A. Hajizadeh and M. Soltani, "Protection in DC microgrids: a comparative review," in IET Smart Grid, vol. 1, no. 3, pp. 66-75, 10 2018.

[17] S. Bernet, T. Matsuo, and T. A. Lipo, “A matrix converter using reverse blocking NPT-IGBTs and optimized pulse patterns,” in Proc. 27th Annu. IEEE Power Electronics Specialists Conf., 1996, pp. 107–113.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 156-166. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2848

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

EXPERIMENTAL METHODS AND A/B TESTING IN FINTECH: APPLIED GUIDE IN PYTHON

Alexander Haro-Sarango*[1], Julian Coronel-Reyes[2], Johnny Triviño-Sanchez[3], Ibeth Molina-Arcos[4], Gerardo Villarreal-Terán[5] and Gladys Proaño-Altamirano[6]

[1, 4, 5, 6]Instituto Superior Tecnológico España, Ambato, Ecuador.
[2, 3]Universidad Agraria del Ecuador, Guayaquil, Ecuador.

[1]https://orcid.org/0000-0001-7398-2760, [2]https://orcid.org/0000-0002-7883-5388, [3]https://orcid.org/0009-0008-2151-4867
[4]https://orcid.org/0000-0001-9650-1317, [5]https://orcid.org/0009-0003-1384-5184, [6]https://orcid.org/0000-0001-6809-7687

Email: *alexander.haro@iste.edu.ec

ARTICLE INFO

Article History
Received: October 23, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Experimental methods,
A/B testing,
FinTech,
Python,
user conversion.

ABSTRACT

This paper provides an applied guide to the use of experimental methods, particularly A/B testing, in the FinTech sector, leveraging Python-based tools. It examines how these techniques validate business hypotheses, optimize interfaces, personalize financial services, and enhance critical metrics such as conversion and retention. Using a rigorous experimental design, the study evaluates treatment effects on conversion rates, deposits, and user activity through t tests, z tests, bootstrap resampling, and regression models. Results confirm a significant uplift in conversion rates but show no substantial effects on deposits or post-conversion engagement, thereby delimiting the scope of impact. Heterogeneous effects across age segments are identified, with strategic implications for targeted personalization and risk management. The article highlights the importance of incorporating ethical safeguards, transparency, and regulatory compliance when experimenting with sensitive financial data, and proposes a replicable, scalable methodological framework for researchers and practitioners.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Over the past decade, the FinTech sector has undergone a significant transformation driven by advances in artificial intelligence, data analytics, and experimental methodologies. The convergence of digital financial services and information technologies has given rise to a new generation of financial products that are more accessible, personalized, and efficient [1]. In this context, the capacity of FinTech firms to make decisions grounded in empirical evidence has become a key source of competitive advantage [2]. Experimental methods and A/B testing in particular have emerged as fundamental tools to validate business hypotheses, optimize user interfaces, personalize financial services, and improve key metrics such as customer retention and customer lifetime value (CLV). A/B testing enables the comparison of two or more variants of a digital product such as a web interface, a personalized interest rate, or a marketing campaign with the aim of determining which one delivers superior performance with respect to a predefined set of metrics[3], [4].

Originating in the domain of classical experimental design, this methodology has been adopted and adapted by high-growth technology companies such as Amazon, Netflix, and Google, and it now constitutes standard practice in the FinTech environment [5]. Nevertheless, the rigorous implementation of A/B testing in financial settings presents substantial challenges: selection bias, interference between users, issues of external validity, and ethical dilemmas associated with the manipulation of financial products [6], [7]. The situation becomes even more complex when experiments are conducted in real time with highly sensitive data, where errors can entail economic and regulatory consequences [8]. In addition, the explosion of data and the adoption of microservice architectures demand that these experiments be automated, reproducible, and scalable a requirement that has been facilitated by modern Python tools such as statsmodels and SciPy, as well as specialized libraries like Evidently, Optimizely, and PyAB. The integration of these technologies has enabled both startups and established banks to implement continuous experimentation systems [9], [10].

This article provides an applied guide to implementing experimental methods and A/B testing in the FinTech context using Python, with the objective of bridging the gap between statistical theory and its practical execution in production systems. The guide is grounded

in recent studies, industry best practices, and real-world use cases in the digital financial industry. It also discusses extensions such as multivariate testing, Bayesian testing, and adaptive experimentation, offering a comprehensive and up-to-date perspective for researchers, developers, and data scientists in FinTech.

II. LITERATURE REVIEW

A/B testing, also known as a randomized controlled experiment, is a widely used technique in applied research to establish causal inferences between a manipulated variable and an observed outcome. Its rise in digital environments is attributable to platforms' ability to randomly assign treatments to large user populations and to record resulting behavior with precision [3]. By directly comparing two variants (A and B), the approach assesses which one yields superior performance on predefined metrics such as conversion rate, user retention, or profitability. On platforms such as Amazon, Google, or Netflix, this technique has been fundamental to the continuous improvement of interfaces and personalized services. [3] and [8] describe how controlled experimentation becomes an operating system for innovation when integrated with automated tooling and rigorous statistical principles. These advances have been transferred to the FinTech domain, where precise measurement of the impact of algorithmic changes, pricing strategies, or design interventions is essential.

The FinTech sector presents particularities that complicate the application of A/B testing: regulatory compliance, the sensitivity of financial data, high interdependence among users, and the need for explainability in automated decisions [9]. and [10].notes that A/B tests can be useful for evaluating alternative credit-scoring tools, financial inclusion strategies, or digital identity validation mechanisms. Nevertheless, such experiments must be conducted under ethical oversight, especially when treatments affect sensitive decisions such as credit approval, assignment of credit limits, or insurance personalization. [11-13] caution that, although randomized experiments are methodologically robust, they can generate unintended consequences if interference effects among users or historical algorithmic biases are not addressed. In this regard, [14], [15] propose the design of experiments on social networks using cluster randomization techniques that control spillovers among connected nodes a particularly useful practice in FinTech applications based on reputation systems or peer-to-peer lending.

Beyond the classical hypothesis-testing approach based on p-values, authors such as [16] and [17] have proposed Bayesian frameworks for experimentation that allow for a more intuitive interpretation of results and the dynamic updating of evidence as data are collected. This is especially relevant in FinTech, where contexts can change rapidly, and decisions must adapt to emerging user behaviors. Bayesian methods yield posterior distributions for the parameters of interest, thereby facilitating probabilistic decision-making. [17] introduces the multi-armed bandit approach, which dynamically adapts treatment allocation to maximize expected value, thus reducing exposure to ineffective alternatives. [18] also highlights the role of Bayesian A/B testing as an accelerator in product-validation processes. These approaches are ideally suited to financial contexts in which risk is inherent to decision-making, and it is necessary not only to identify the superior alternative but also to quantify uncertainty and minimize the associated losses.

From a computational standpoint, Python has become the dominant language for experimentation in data science due to its flexibility, active community, and rich library ecosystem. Tools such as statsmodels, SciPy, PyAB, Evidently, and Bayesian PyAB enable the implementation of A/B tests, analysis of key metrics, automated visualizations, and real-time monitoring of experiment validity [19]. Additionally, commercial platforms like Optimizely and Google Optimize offer APIs that integrate with Python workflows, facilitating the execution of experiments within production data pipelines. [20], [21] propose microservice-based continuous experimentation architectures, which are especially relevant for FinTechs seeking to scale products across multiple market segments. Tools such as MLflow, Dagster, and Airflow make it possible to combine experimental design with DevOps and MLOps practices, thereby facilitating reproducibility, auditing, and traceability crucial aspects for compliance with financial regulations such as GDPR, PSD2, and banking-supervision requirements.

The use of A/B testing in FinTech raises important ethical challenges. Experimental decisions can affect a user's eligibility for credit, insurance, or investment services; consequently, transparency and fairness become essential pillars [6], [22]. Recent literature emphasizes the need to audit algorithms and to ensure that tests do not reinforce existing biases or generate indirect discrimination. [22] introduce the concept of "algorithmic audits" to evaluate the social impact of automated systems, while [23] analyze mechanisms for explainability and accountability in financial algorithmic contexts. These perspectives are fundamental when implementing experimental methods that affect sensitive variables such as interest rates, credit limits, or approval rates. Therefore, the specialized literature broadly supports the use of A/B testing in FinTech, provided that robust methodological safeguards, appropriate technological tools, and ethical principles aligned with fairness and user protection are adopted.

III. MATERIALS AND METHODS

From an anonymized dataset of 10,000 customers of a Segment-1 cooperative savings and credit institution in Ecuador, the study adopts a classical A/B randomized experiment to estimate the causal effect of a digital Fintech intervention (e.g., a redesigned onboarding flow, a promotional banner, or a new deposit screen). Each observation corresponds to a unique user exposed once during the study window. Assignment to control A or treatment B was performed via simple randomization with probability $P(B) = 0.5$, without replacement and at the user level. Under protocol compliance, the treatment indicator $D_i = \mathbf{1}\{i \in B\}$ is independent of the potential outcomes $(Y_i(0), Y_i(1))$. This independence guarantees identifiability of the Average Treatment Effect (ATE) as

$$\tau = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0]. \quad (1)$$

The primary outcome is the binary conversion $Y_i^{\text{conv}} \in \{0,1\}$, which flags whether the user completes the target action (e.g., onboarding or first deposit). Secondary outcomes capture usage intensity: deposited amount among converters $Y_i^{\text{dep}} \geq 0$, number of logins login_count_i, and active days days_active_i.

Pre-treatment covariates used for balance checks and precision gains include age_i, self-reported $income_i$, and historical activity metrics prior to the intervention. Duplicates were removed, missingness was documented, and monetary variables were winsorized at the

top 1% to mitigate outliers while preserving mean interpretability. Before effect estimation, balance between groups is validated through mean-comparison tests on pre-treatment covariates. For each continuous variable $X \in \{age, income,$ login_count, days_active$\}$ we apply a two-sided Welch t-test:

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}, \tag{2}$$

And report the associated p-value. Systematic non-significance (e.g., $p > 0.05$) supports proper randomization and thus the experiment's internal validity. Overlaid histograms by group are inspected to rule out substantial imbalances not captured by the mean. The main hypothesis tests whether conversion rates differ between A and B. Let $p_A = \mathbb{E}[Y^{\text{conv}} \mid D = 0]$ and $p_B = \mathbb{E}[Y^{\text{conv}} \mid D = 1]$. The two-proportion z-statistic is

$$\hat{z} = \frac{p_B - p_A}{\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}, \qquad \hat{p} = \frac{n_A p_A + n_B p_B}{n_A + n_B} \quad , \tag{3}$$

Under $H_0: p_B = p_A$. We report point estimates, a 95% confidence interval (CI), and a two-sided p-value. Business magnitude is also communicated as the relative lift,

$$\text{lift} = \frac{p_B - p_A}{p_A}. \tag{4}$$

For the deposited amount among converters, we estimate a conditional mean effect by comparing Y^{dep} in the sub-sample $\{i: Y_i^{\text{conv}} = 1\}$. Given its continuous and skewed nature, we contrast mean differences via Welch's t-test and accompany the analysis with boxplots to visualize median and dispersion; when appropriate, robustness is checked with a 10% trimmed mean or a bootstrap difference in medians. To characterize treatment heterogeneity, the sample is stratified ex-post by age bands $S \in \{< 30,\ 30\text{–}50,\ > 50\}$ using binning on $[18{,}30{,}50{,}70)$. For each segment we estimate

$$\hat{\tau}_S = \mathbb{E}[Y^{\text{conv}} \mid D = 1, S] - \mathbb{E}[Y^{\text{conv}} \mid D = 0, S \tag{5}$$

These analyses are descriptive and interpreted cautiously to avoid multiple-comparison pitfalls; when a segment shows signal, we complement with bootstrap CIs and recommend future ex-ante stratification in the sampling design. Uncertainty around the ATE in conversion is further quantified via nonparametric bootstrap with independent resampling within A and B. For iterations $b = 1, \ldots, B$ (with $B = 5000$), we draw with replacement $A^{(b)}$ and $B^{(b)}$ matching original sizes and compute $\hat{\tau}^{(b)} = \bar{Y}_{B^{(b)}} - \bar{Y}_{A^{(b)}}$. The 95% CI is given by $[q_{2.5},\ q_{97.5}]$ of the empirical distribution $\{\hat{\tau}^{(b)}\}$, providing a normal-assumption-free interval particularly useful with low rates or unequal sizes. To gain precision and control residual noise, we estimate a linear probability model of conversion on treatment and pre-treatment covariates:

$$Y_i^{\text{conv}} = \beta_0 + \beta_1 D_i + \gamma_1\, age_i + \gamma_2\, income_i + \gamma_3\, \text{login_count}_i + \gamma_4\, \text{days_active}_i + \varepsilon_i. \tag{6}$$

Under randomization and protocol compliance, the causal interpretation of β_1 coincides with the ATE, and we report OLS estimates with heteroskedasticity-robust (HC3) standard errors. As a sensitivity check, we also fit a reduced model

$$Y_i^{\text{conv}} = \beta_0 + \beta_1 D_i + \eta_i, \tag{7}$$

To verify that adding controls improves efficiency without altering the sign or order of magnitude. For continuous outcomes (Y^{dep}, login_count, days_active) we apply analogous OLS specifications; if heavy tails persist, we test log-transformations on $(Y^{\text{dep}} + c)$. Placebo and bias validations include tests on variables that should not respond to treatment, such as pre-intervention income. A Welch t-test on $income$ between A and B is expected to be non-significant; otherwise, we diagnose potential randomization issues or attrition. We additionally inspect missingness rates by column and the distribution of sample sizes by group to rule out differential measurement bias. Inference adopts $\alpha = 0.05$ with two-sided tests, and Holm–Bonferroni adjustments are noted when evaluating multiple secondary outcomes. For ex-ante statistical power, we use the standard approximation for a difference in proportions with 1:1 allocation. Given significance α and power $1 - \beta$, the Minimum Detectable Effect (MDE) around a baseline rate p satisfies

$$\text{MDE} = \left(z_{1-\alpha/2} + z_{1-\beta}\right)\sqrt{\frac{4\,p(1-p)}{n}}, \tag{8}$$

Where n is the total sample size. This quantifies whether $n = 10{,}000$ is adequate for the business-relevant lift. Although the analysis is ex-post, reporting the MDE contextualizes the sensitivity of findings. The workflow is implemented reproducibly in Python: pandas for data handling, scipy.stats for t- and z-tests, statsmodels for OLS with robust errors, and plotly for interactive figures. The pipeline proceeds through reading and cleaning, balance verification, the primary conversion test, analysis of financial and activity metrics, age-based heterogeneity, bootstrap uncertainty quantification, and robustness regressions, followed by placebo and experiment-quality checks. Code is organized in self-contained cells to facilitate auditability and replication. Regarding validity, randomization and balance checks support internal validity; external validity is restricted to the cooperative's clientele and time period analyzed. The intervention is purely digital and does not alter contractual conditions or fees, so participant risk is minimal.

Privacy and data-security principles are observed through irreversible anonymization, attribute minimization, and exclusive use for product-evaluation purposes, consistent with the local personal-data protection framework. Finally, interpretation is aligned with

Fintech decision-making: if the p-value is below 0.05 and the ATE CI excludes zero with business-relevant magnitude, the *B* variant is recommended; if effects are null or uncertain, we favor the lower-cost alternative or iterate the design incorporating segment learning with ex-ante stratification. This protocol offers an applied, mathematically grounded, and operational guide for product teams in cooperative financial institutions, scalable to continuous experimentation cycles with robust causal rigor.

IV. RESULTS AND DISCUSSIONS

It is the logical presentation of the results that demonstrate the true contribution of the research and also justify the conclusion. The parts indicated above, Introduction, Material and Methods serve to explain how the results are obtained. Use tables and figures for your explanation. The results must be written in the past, in a brief and simple way. Avoid redundancy. Analyzing and interpreting the results is part of the discussions, their importance, achievements and limitations, highlighting the innovative aspects of the practical applications of the study and the conclusions derived from them, delimiting unresolved issues. If necessary, recommendations can be proposed. In this first stage of the experimental analysis, the aim is to verify that the random assignment of users to groups A (control) and B (treatment) was effective that is, that both groups are comparable prior to any causal analysis.

To this end, several observable variables are examined: age, income, number of logins (login_count), and days active on the platform (days_active). Validation is performed by comparing the means of these variables across groups using an independent-samples t test. This procedure assesses whether there are statistically significant differences in the means of the variables considered. The p-value obtained in each t test indicates the probability of observing a difference as large as the one found if, in fact, no difference exists in the population. In this case, the p-values were 0.2378 for age, 0.9431 for income, 0.4110 for login count, and 0.8552 for days active. All of these values are well above the conventional 0.05 threshold, so we conclude that there are no significant differences between groups. The analysis is complemented by a descriptive summary of the means for each variable in both groups.

For example, the average age in group A was 34.94 years, whereas in group B it was 34.72 years. Similarly, average income was virtually identical: 30,052 for A and 30,066 for B. The average number of logins was 5.01 in A versus 4.98 in B, and days active were 14.98 and 14.95, respectively. These figures indicate that both groups exhibit nearly identical characteristics, reinforcing the validity of the randomization. To visualize these results, histograms generated with the Plotly library were used. These plots display how each variable is distributed within each group and allow any notable overlaps or differences to be observed. The distributions for age, income, logins, and days active show a high degree of visual similarity between groups A and B, confirming the statistical findings (see Table 1).

Table 1: Step 1 – Validation of balance between groups.

Variable	p-value	Mean Group A	Mean Group B
age	0.2378	34.94	34.72
income	0.9431	30,051.88	30,065.77
login_count	0.4110	5.01	4.98
days_active	0.8552	14.98	14.95

Source: Authors, (2026).

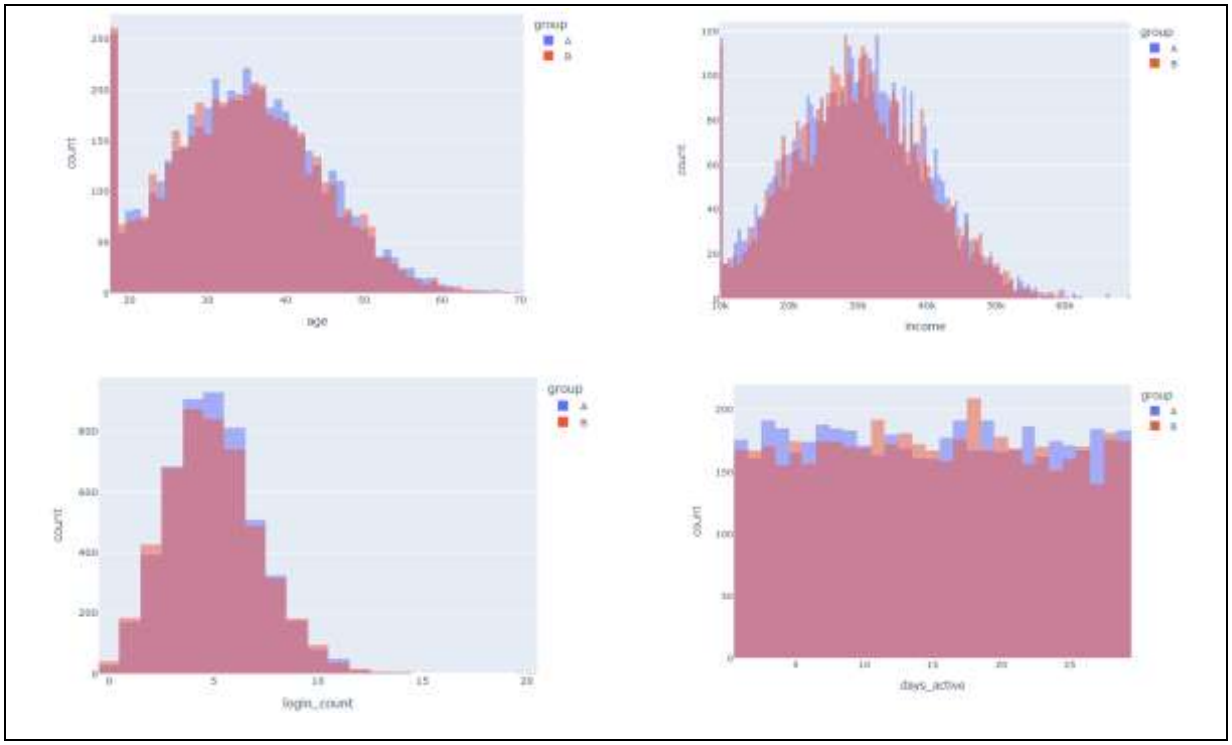


Figure 1: Step 1 – Distribution among groups.
Source: Authors, (2026).

In the second step of the analysis, we evaluate whether the treatment applied to group B had a significant effect on conversion; that is, whether the proportion of users who performed the desired action (e.g., opening an account or completing a registration) is higher relative to group A. To this end, we compute the conversion rate in each group and then apply a formal statistical test. In this case, a z test for the difference in proportions specifically designed to compare the success rates of two independent groups was used. The results show that the conversion rate in group A was 0.0985, whereas in group B it was 0.1509. This difference represents a relative increase of more than 50%. To determine whether this difference is statistically significant, we calculated the z statistic, which was −7.9389. The associated p-value was less than 0.0001, indicating that the probability of observing so large a difference by chance alone, if no true effect existed, is virtually zero. Consequently, we reject the null hypothesis of equal proportions and conclude that the treatment had a significant effect on conversion. The observed difference is also depicted with a bar chart, which clearly shows that the height of the bar corresponding to group B is substantially greater than that of group A. This provides a clear and intuitive visualization of the treatment's positive effect. In sum, this analysis offers strong statistical evidence that the intervention applied to group B significantly increased the user conversion rate (see Table 2).

Table 2: Step 2: Conversion analysis and hypothesis testing.

Metric	Group A	Group B	Test Statistic	p-value
Conversion Rate	0.0985	0.1509	z = -7.9389	0.0000

Source: Authors, (2026).

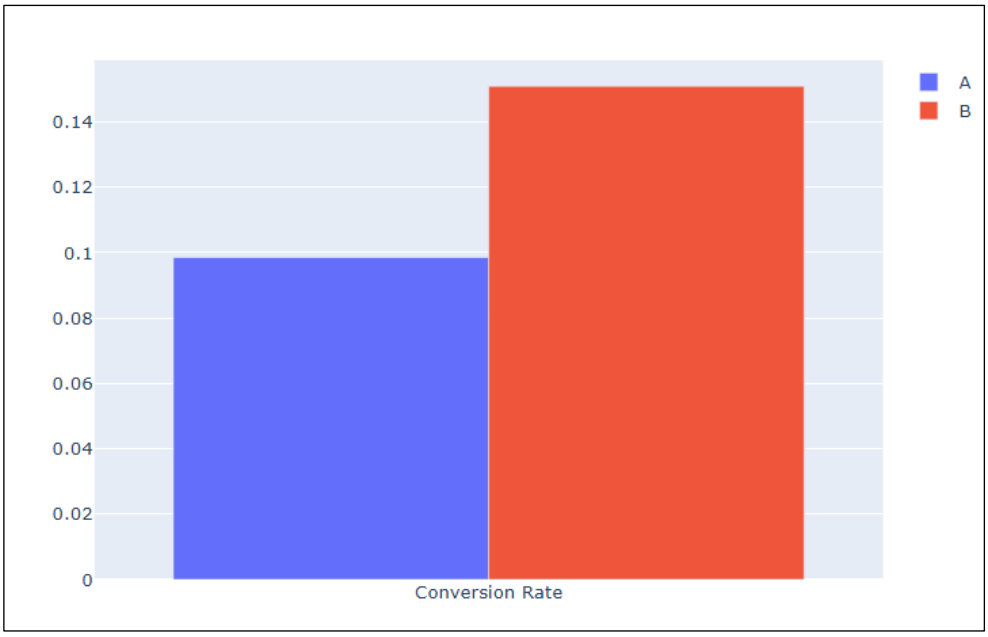


Figure 2: Step 2: Conversion rate by group.
Source: Authors, (2026).

In this third step, we analyze whether the treatment affected not only conversion but also users' subsequent financial and operational behavior specifically, the amount deposited, the number of times they logged into the platform, and the days they remained active. These indicators are crucial for assessing conversion quality: it is not sufficient to convert more users if they subsequently do not interact or generate revenue. To begin, an independent-samples t test was applied to compare the deposited amount between groups A and B, considering only users who actually converted. The result was a t value of −0.4683 and a p value of 0.6397, indicating no statistically significant difference between the two groups in terms of deposited amount. Although the group B distribution appears visually to have a slightly higher median, variability in the data and overlap between distributions render this difference inconclusive from a statistical standpoint.

Boxplots make this pattern clear. In the deposited-amount plot, both the median and the dispersion are similar across groups. A handful of outliers appear on both sides as is to be expected with financial data but there is no evidence of a systematic treatment effect. With respect to the number of logins (login_count) and days active (days_active), the boxplots likewise show very similar distributions for groups A and B, both in terms of median and interquartile range. This suggests that post-conversion behavior, in terms of platform use, was not altered by the treatment. Although group B achieved a significantly higher conversion rate, the users who converted in that group did not deposit more money nor were they more active than those in group A. This is an important finding because it helps distinguish between superficial conversion and conversion that yields real business value (see Table 3).

Table 3: Step 3: Analysis of Financial Metrics (Amount and Activity).

Metric	Test Statistic	p-value
Deposited Amount	t = -0.4683	0.6397

Source: Authors, (2026).

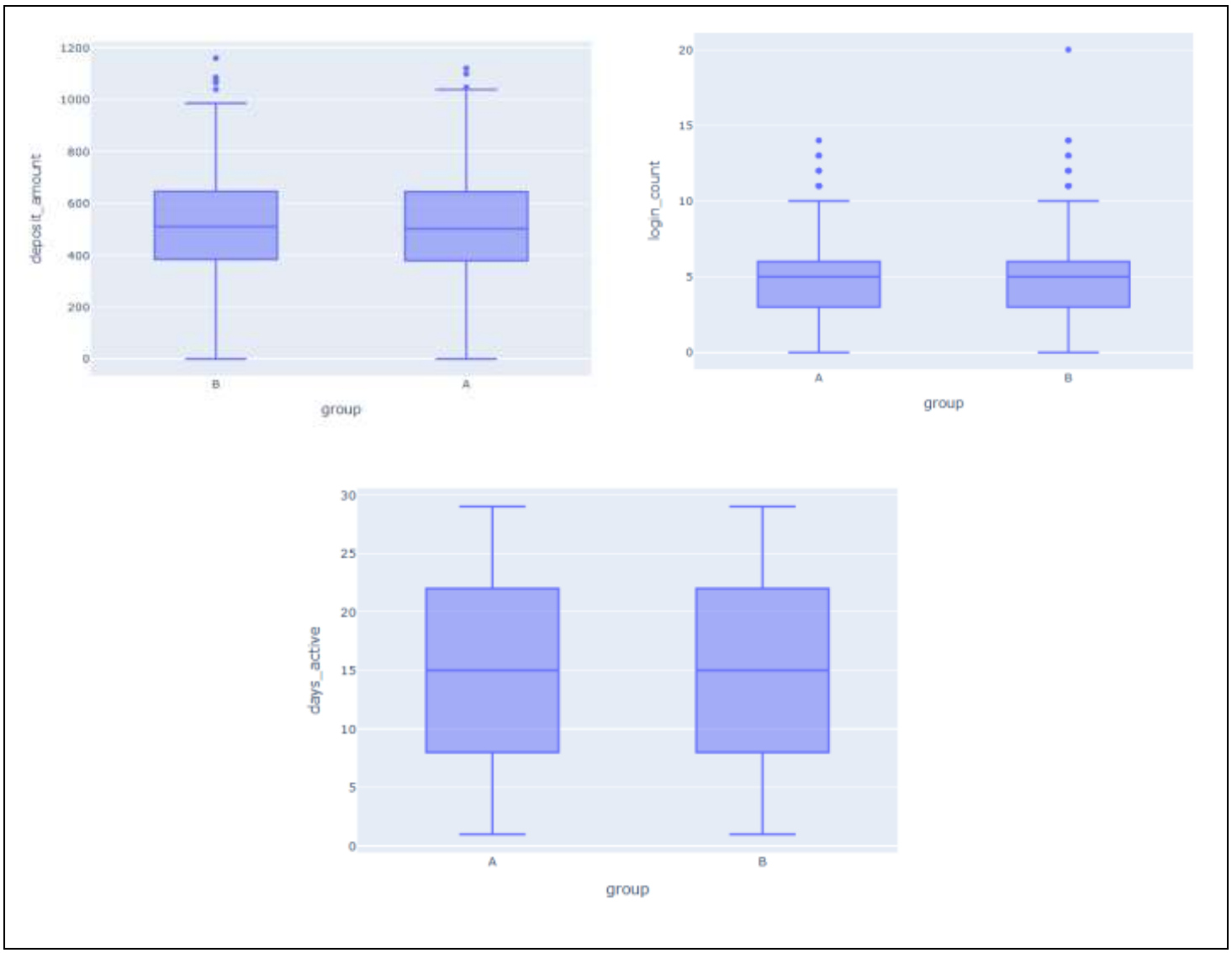


Figure 3: Step 3: Conversion rate by group.
Source: Authors, (2026).

At this stage (see Table 4), we investigate whether the treatment's impact varies by users' age segment, thereby identifying potential heterogeneous effects. To this end, the age variable was partitioned into three brackets: under 30, 30–50, and over 50. We then computed the conversion rate for each age group, disaggregated by experimental arm (A or B). This segmentation makes it possible to detect whether the treatment was more effective for certain demographic profiles. The results show that group B consistently outperformed group A across all three age segments. Among users under 30, the conversion rate was 9.4% in group A and 16.1% in group B, an absolute difference of 6.8 percentage points. In the 30–50 segment, the rates were 10.0% for A and 14.5% for B, a 4.5-point difference. Among users over 50, group B reached a conversion rate of 16.2%, compared with 8.9% for group A the largest observed difference (7.3 percentage points).

A grouped bar chart reinforces this reading. In every segment, the bars corresponding to group B are notably higher than those for group A, indicating that the treatment's positive effect holds across age ranges. However, the impact is slightly larger at the extremes (<30 and >50), suggesting that these groups may be more sensitive to the change introduced by the experiment. This segmented analysis is valuable because it reveals that, although the treatment is beneficial overall, its magnitude varies across user subgroups. Such information can be used to design more personalized campaigns or to allocate the treatment more efficiently in future deployments. It also rules out the possibility that the observed effect on overall conversion is driven by a single segment, which adds robustness to the conclusions.

Table 4: Step 4: Treatment heterogeneity analysis (segmented).

Segment	Conversion A	Conversion B	Difference
< 30	0.094	0.161	0.068
30–50	0.100	0.145	0.045
> 50	0.089	0.162	0.073

Source: Authors, (2026).

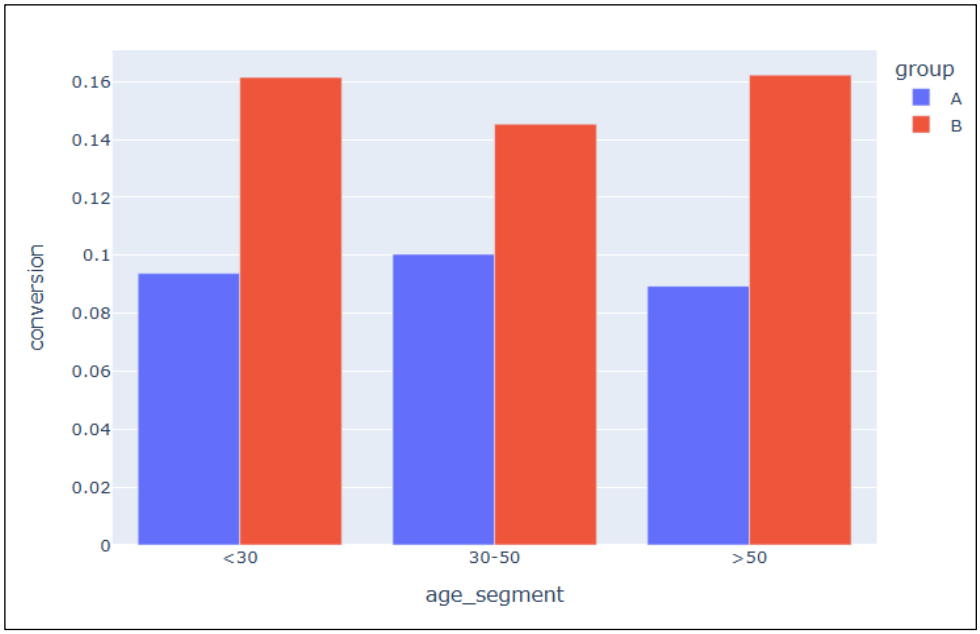


Figure 4: Step 4: Distribution in bar chart Treatment heterogeneity analysis (segmented).
Source: Authors, (2026).

In this fifth step, we apply a resampling technique bootstrap to robustly estimate the average treatment effect (ATE) on conversion together with a confidence interval. Unlike traditional parametric tests, bootstrap does not assume normality or homoscedasticity; instead, it draws many random samples with replacement from the original data and computes the statistic of interest at each iteration. Here, 5,000 iterations were performed, computing the difference in conversion rates between group B and group A for each resampled dataset. This procedure yields an empirical distribution of the ATE, from which a 95% confidence interval can be obtained by extracting the 2.5th and 97.5th percentiles. The resulting mean ATE was 0.0524 that is, an absolute difference of 5.24 percentage points in favor of group B. The confidence interval ranged from 0.0391 to 0.0652, indicating that, with 95% confidence, the true treatment effect lies within this interval. The accompanying plot depicts the distribution of conversion-rate differences across all simulations. The curve is approximately normal, centered around the mean value (0.0524), with red dashed lines marking the confidence-interval bounds. This visualization shows that virtually all simulations yielded a positive effect, reinforcing confidence in the robustness of the result (see Table 5).

Table 5: Step 5: Bootstrap-based confidence intervals for the treatment effect.

Metric	**Estimate**	**95% CI (Lower – Upper)**
ATE (Bootstrap)	0.0524	0.0391 – 0.0652

Source: Authors, (2026).

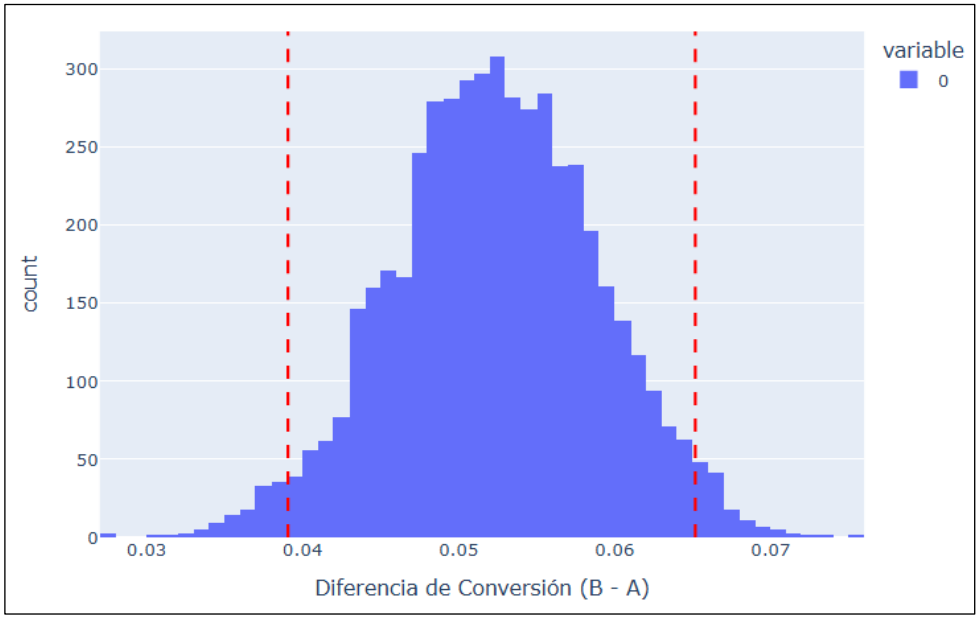


Figure 5: Step 5: Histogram of Bootstrap-based confidence intervals for the treatment effect.
Source: Authors, (2026).

In this step (see Table 6), an ordinary least squares (OLS) regression model is used to estimate the causal effect of the treatment on conversion while controlling for additional variables. Although conversion is a binary variable (0 or 1), OLS can be employed as an initial approximation to the average treatment effect, and it also allows inspection of the influence of other explanatory variables. In this model, the dependent variable is conversion, and the predictors are the assigned group (group), age, income, number of logins, and days active. The most relevant coefficient is that associated with group B (C(group)[T.B]), which has an estimated value of 0.0523 with a standard error of 0.0066. This implies that, holding the other variables constant, belonging to group B increases the probability of conversion by approximately 5.23 percentage points—a result that is highly consistent with the findings in the previous steps.

In addition, the p-value associated with this coefficient is less than 0.0001, confirming that the effect is statistically significant. This p-value measures the probability that so large a difference would occur by chance if group B truly had no effect. As for the remaining variables, none is statistically significant: age has a negative coefficient but a p-value of 0.207; income has a p-value of 0.337; the number of logins is irrelevant (p = 0.383); and days active likewise shows no effect (p = 0.278). This indicates that, within this model, the only factor clearly associated with conversion is assignment to the treatment group. The model's R-squared is very low (0.007), which is common in models with binary outcomes and reflects that most individual variability is not explained by the included covariates. However, the objective here is not to predict conversion but to estimate the causal effect of the treatment, which is clearly obtained.

Table 6: Step 6: Estimating the causal effect with regression (OLS).

Field			**Value**			
Dep. Variable			conversion			
Model			OLS			
Method			Least Squares			
Date			Sat, 19 Jul 2025			
Time			00:35:36			
No. Observations			10000			
Df Residuals			9994			
Df Model			5			
Covariance Type			nonrobust			
R-squared			0.007			
Adj. R-squared			0.006			
F-statistic			13.57			
Prob (F-statistic)			3.21e-13			
Log-Likelihood			-3066.6			
AIC			6145.0			
BIC			6188.0			
Variable	**coef**	**std err**	**t**	**P>|t|**	**0.025**	**0.975**
Intercept	0.1040	0.019	5.541	0.000	0.067	0.141
C(group)[T.B]	0.0523	0.007	7.942	0.000	0.039	0.065
age	-0.0004	0.000	-1.262	0.207	-0.001	0.000
income	3.247e-07	3.38e-07	0.959	0.337	-3.39e-07	9.88e-07
login_count	-0.0013	0.001	-0.873	0.383	-0.004	0.002
days_active	0.0004	0.000	1.086	0.278	-0.000	0.001
Statistic			**Value**			
Omnibus			4113.569			
Prob(Omnibus)			0.000			
Skew			2.254			
Kurtosis			6.143			
Durbin–Watson			1.992			
Jarque–Bera (JB)			12586.243			
Prob(JB)			0.000			
Cond. No.			1.81e+05			

Source: Authors, (2026).

In this seventh step (see Table 7), the goal is to validate the robustness of the experiment, ensuring that the results obtained are not artifacts of methodological errors, hidden biases, or spurious effects. To this end, three complementary strategies are applied: a placebo test, a review of experimental integrity, and a sensitivity analysis with simplified models. The placebo test consists of applying the same type of statistical comparison used to measure the treatment effect, but to a variable that should not be influenced by the treatment. In this case, income was chosen as the placebo variable, since there is no reason to expect that belonging to group A or B would alter users' income. A t test was conducted to compare income across groups, yielding a t value of −0.0714 and a p value of 0.9431. Such a high p-value indicates that there is no significant difference between groups, which validates that the experimental design is not generating systematic false positives. Next, a quality review of the data and assignment was carried out. No missing values were found in any key variable of the experiment, indicating that there was no dropout or data loss in the process.

The only exception was the age_segment column used solely for the segmented analysis which contains 517 null values due to users whose ages fell outside the defined range. Furthermore, the distribution of users across groups was very close to 50%, with 50.76% in group A and 49.24% in group B, confirming that random assignment was executed correctly.

Table 7: Step 7: Robustness checks (placebo, integrity review, and sensitivity analysis).

Variable	**Test Statistic**	**p-value**
Income	t = -0.0714	0.9431
Column	**Missing Values**	
user_id	0	
group	0	
age	0	
income	0	
conversion	0	
deposit_amount	0	
login_count	0	
days_active	0	
age_segment	517	
Group	**Proportion**	
A	0.5076	
B	0.4924	

Source: Authors, (2026).

Subsequently, a sensitivity verification (see Table 8) was conducted using a reduced regression model that included only two variables: group and age. The objective is to assess whether the coefficient for group B remains stable when other covariates are removed. The results show that the coefficient for group B remains exactly 0.0523, with a p-value less than 0.0001, indicating that the treatment effect does not depend on the inclusion of additional variables in the model. This stability reinforces confidence that the observed effect is robust and not conditioned by model specification.

Table 8: Step 7 (part 2) – Sensitivity check with partial regression.

Field	**Value**
Dep. Variable	conversion
Model	OLS
Method	Least Squares
Date	Sat, 19 Jul 2025
Time	00:39:49
No. Observations	10000
Df Residuals	9997
Df Model	2
Covariance Type	nonrobust
R-squared	0.006
Adj. R-squared	0.006
F-statistic	32.49
Prob (F-statistic)	8.64e-15
Log-Likelihood	-3068.0
AIC	6142.0
BIC	6164.0

Variable	**coef**	**std err**	**t**	**P>|t|**	**0.025**	**0.975**
Intercept	0.1135	0.013	8.806	0.000	0.088	0.139
C(group)[T.B]	0.0523	0.007	7.948	0.000	0.039	0.065
age	-0.0004	0.000	-1.248	0.212	-0.001	0.000

Statistic	**Value**
Omnibus	4115.683
Prob(Omnibus)	0.000
Skew	2.255
Kurtosis	6.145
Durbin–Watson	1.992
Jarque–Bera (JB)	12599.341
Prob(JB)	0.000
Cond. No.	143.0

Source: Authors, (2026).

V. CONCLUSIONS

The experimental program provides convergent evidence that the treatment materially improves user conversion while preserving internal validity. Random assignment yielded well-balanced groups across observable covariates (age, income, login_count, days_active), and classical tests failed to detect any pre-treatment imbalances. In the primary outcome, the treatment produced a large and statistically precise lift in conversion rising from 9.85% in the control arm to 15.09% in the treated arm corroborated by a z-test with $p<0.0001$. Bootstrap estimation further quantified this effect with a mean average treatment effect (ATE) of 5.24 percentage points and a narrow 95% confidence interval [3.91%, 6.52%], indicating that the uplift is not an artifact of parametric assumptions and is robust to resampling uncertainty. Causal estimation with a linear probability model (OLS) that controlled for user demographics and baseline activity yielded a treatment coefficient of 0.0523 (SE=0.0066; $p<0.0001$), essentially identical to the nonparametric bootstrap estimate. Taken together, the randomized comparison, the resampling analysis, and the regression adjustment converge on the same quantitative conclusion: the intervention consistently increases the probability of conversion by about five percentage points.

Robustness checks placebo testing on income, integrity audits of missingness and assignment ratios, and a reduced-form sensitivity model found no evidence of hidden bias, data leakage, or specification dependence, reinforcing the credibility of the causal claim. At the same time, secondary outcomes delineate the boundary of the treatment's impact. Among converters, the intervention did not increase deposited amounts nor post-conversion engagement (logins and active days); distributions and inferential tests show no statistically meaningful differences between arms on these metrics. This pattern implies that the treatment primarily improves top-of-funnel behavior (initial conversion) without yet translating into deeper behavioral or financial value. In managerial terms, the intervention appears to drive "shallow" conversions rather than high-value activations, a distinction that is crucial for revenue forecasting, lifetime-value modeling, and budgeting of acquisition spend. Heterogeneity analyses by age segment show that the uplift is broad-based but not uniform: effects are directionally larger in users under 30 and over 50, with mid-age users showing a smaller yet still positive gain.

This profile suggests immediately actionable targeting policies. In the short run, firms could prioritize deployment in the more responsive segments to maximize near-term ROI while designing complementary nudges aimed at upgrading mid-segment users from initial conversion to meaningful activation. In the medium term, these findings motivate segment-specific experimentation (e.g., multivariate tests tuned to frictions and value propositions salient for each cohort). These results carry two strategic implications. First, acquisition and activation should be treated as distinct optimization problems: the tested treatment is a strong lever for conversion, but it should be paired with post-onboarding interventions (e.g., personalized prompts, financial education micro-journeys, product bundling, or rate-structure trials) to unlock deposit growth and durable engagement. Second, evaluation should migrate from single-metric hypothesis tests to multi-objective experimental design that jointly tracks conversion, early-value proxies (first-week deposits, feature adoption), and forward-looking business KPIs (predicted CLV, retention hazard).

Doing so will align experimental success criteria with financial materiality. Methodologically, the study's triangulation across classical tests, bootstrap inference, and covariate-adjusted models is a strength, as is the suite of robustness diagnostics. Future extensions could corroborate the estimates with generalized linear models (logit/probit), apply Bayesian A/B frameworks for continuous monitoring and decision thresholds, and incorporate adaptive allocation (multi-armed bandits) to improve sample efficiency while respecting guardrails for fairness and regulatory compliance. Because the treatment does not yet shift value-creation metrics, uplift modeling at the user level (treatment effect heterogeneity on deposits/engagement) and experiment-within-experiment designs (sequential interventions after conversion) are natural next steps. Finally, given the financial context, governance remains central. Any scale-up should include pre-registered analysis plans, segment-level fairness audits, and continuous monitoring for interference and spillovers (especially where social or referral effects may arise). By coupling disciplined experimentation with ethical safeguards and value-focused KPIs, organizations can translate the demonstrated conversion uplift into sustained, equitable, and auditable business impact.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Alexander Haro-Sarango, Julian Coronel-Reyes, Johnny Triviño-Sanchez, Ibeth Molina-Arcos, Gerardo Villarreal-Terán and Gladys Proaño-Altamirano.
Methodology: Alexander Haro-Sarango, Julian Coronel-Reyes.
Investigation: Alexander Haro-Sarango, Julian Coronel-Reyes, Johnny Triviño-Sanchez.
Discussion of results: Alexander Haro-Sarango, Julian Coronel-Reyes, Johnny Triviño-Sanchez, Ibeth Molina-Arcos, Gerardo Villarreal-Terán and Gladys Proaño-Altamirano.
Writing – Original Draft: Alexander Haro-Sarango, Julian Coronel-Reyes, Johnny Triviño-Sanchez, Ibeth Molina-Arcos, Gerardo Villarreal-Terán and Gladys Proaño-Altamirano.
Writing – Review and Editing: Alexander Haro-Sarango, Julian Coronel-Reyes.
Resources: Alexander Haro-Sarango, Julian Coronel-Reyes, Johnny Triviño-Sanchez, Ibeth Molina-Arcos.
Supervision: Alexander Haro-Sarango, Julian Coronel-Reyes, Johnny Triviño-Sanchez, Ibeth Molina-Arcos, Gerardo Villarreal-Terán and Gladys Proaño-Altamirano.
Approval of the final text: Alexander Haro-Sarango, Julian Coronel-Reyes, Johnny Triviño-Sanchez, Ibeth Molina-Arcos, Gerardo Villarreal-Terán and Gladys Proaño-Altamirano.

VII. REFERENCES

[1] Gai K, Qiu M, Sun X. A survey on FinTech. J Netw Comput Appl. 2018;103: 262–273. doi:10.1016/j.jnca.2017.10.011

[2] Gomber P, Kauffman RJ, Parker C, Weber BW. On the fintech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services. J Manag Inf Syst. 2018;35: 220–265. doi:10.1080/07421222.2018.1440766

[3] Kohavi R, Tang D, Xu Y. Trustworthy online controlled experiments: A practical guide to a/b testing. 2020. Available: https://books.google.com.ec/books?hl=es&lr=&id=Gu-

CEAAAQBAJ&oi=fnd&pg=PP1&dq=Trustworthy+Online+Controlled+Experiments:+A+Practical+Guide+to+A/B+Testing+%5BInternet%5D.+Cambridge&ots=w57WGTgP5l&sig=OsWBsb-RluhO6B8SMgfL1cj3nsk.

[4] Zou T, Xiong F, Li S, Zhang W. Understanding the determinants of firms' usage of A/B testing: A technology–organization– environment framework. IEEE Trans Eng Manage. 2025;72: 378–400. doi:10.1109/tem.2024.3415500

[5] Malovaná S, Janků J, Hodula M, Green G, Joy M, Von Rüden L, et al. Macroprudential policy and income inequality: The trade-off between crisis prevention and credit redistribution. ijcb.org; 2025 [cited 22 Oct 2025]. Available: https://www.ijcb.org/journal/ijcb25q4a5.pdf

[6] Pearl J. Lord's paradox revisited – (oh Lord! Kumbaya!). J Causal Inference. 2016;4. doi:10.1515/jci-2016-0021

[7] Årnes A, Fjeld MK, Stigum H, Nielsen C, Stubhaug A, Johansen A, et al. Does pain tolerance mediate the effect of physical activity on chronic pain in the general population? The Tromsø Study. Pain. 2024. doi:10.1097/j.pain.0000000000003209

[8] Nudurupati SS, Tebboune S, Garengo P, Daley R, Hardman J. Performance measurement in data intensive organisations: resources and capabilities for decision-making process. Prod Plan Control. 2024;35: 373–393. doi:10.1080/09537287.2022.2084468

[9] Kuai L, Wei H. The role of serendipity in narratives: How serendipitous story promotes product interest. Psychol Mark. 2025;42: 1346–1360. doi:10.1002/mar.22181

[10] Chen X, Lin A, Webber S. "We do not always enjoy surprises": investigating artificial serendipity in an online marketplace context. J Doc. 2025;81: 403–422. doi:10.1108/jd-01-2024-0011

[11] Hashi I. Machine Learning–Driven Fintech Solutions for Credit Scoring and Financial Inclusion in the Gig Economy. International Journal of Innovative Science and Research Technology. 2025;10: 1805–1820. Available: https://www.academia.edu/download/124446866/IJISRT25AUG1023.pdf

[12] Óskarsdóttir M, Bravo C, Sarraute C, Vanthienen J, Baesens B. The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. Appl Soft Comput. 2019;74: 26–39. doi:10.1016/j.asoc.2018.10.004

[13] Alamsyah A, Hafidh AA, Mulya AD. Innovative credit risk assessment: Leveraging social media data for inclusive credit scoring in Indonesia's fintech sector. J Risk Fin Manag. 2025;18: 74. doi:10.3390/jrfm18020074

[14] Ugander J, Karrer B, Backstrom L, Kleinberg J. Graph cluster randomization: network exposure to multiple universes. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM; 2013. doi:10.1145/2487575.2487695

[15] Holtz D, Lobel F, Lobel R, Liskovich I, Aral S. Reducing interference bias in online marketplace experiments using cluster randomization: Evidence from a pricing meta-experiment on Airbnb. Manage Sci. 2024. doi:10.1287/mnsc.2020.01157

[16] Camlin CS, Sheira LA, Kwena ZA, Charlebois ED, Agot K, Moody J, et al. The effect of a social network-based intervention to promote HIV testing and linkage to HIV services among fishermen in Kenya: a cluster-randomised trial. Lancet Glob Health. 2025;13: e669–e678. doi:10.1016/S2214-109X(24)00539-4

[17] Phénix T, Ginestet É, Valdois S, Diard J. Visual attention matters during word recognition: A Bayesian modeling approach. Psychon Bull Rev. 2025;32: 1165–1203. doi:10.3758/s13423-024-02591-4

[18] Taddy M. Business data science. 2019. Available: http://222.254.35.8/handle/TLU/8894

[19] Lin C, Ouyang Z, Wang X, Li H, Huang Z. Preserve integrity in realtime event summarization. ACM Trans Knowl Discov Data. 2021;15: 1–29. doi:10.1145/3442344

[20] Niu W, Huang L, Chen M. Spanning from diagnosticity to serendipity: An empirical investigation of consumer responses to product presentation. Int J Inf Manage. 2021;60: 102362. doi:10.1016/j.ijinfomgt.2021.102362

[21] Bao Z, Zhu Y. Understanding online reviews adoption in social network communities: an extension of the information adoption model. Inf Technol People. 2025;38: 48–69. doi:10.1108/itp-03-2022-0158

[22] Bahiru TK, Sinshaw NT, Moges TH, Singh DK. Auditing and mitigating bias in gender classification algorithms: A data-centric approach. arXiv [cs.CV]. 2025. doi:10.48550/arXiv.2510.17873

[23] Binns R, Van Kleek M, Veale M, Lyngs U, Zhao J, Shadbolt N. "it"s reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM; 2018. doi:10.1145/3173574.3173951

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 167-175. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2849

RESEARCH ARTICLE OPEN ACCESS

EFFECT OF BATTERED PILES ON THE RESPONSE OF PILED RAFT AND PILE GROUP FOUNDATIONS: A NUMERICAL STUDY

Toudert Bougrires[*1], Sadok Benmebarek[2], Naïma Benmebarek[3]

[1, 2, 3] MN2I2S Laboratory, Mohamed Khider University of Biskra, Biskra, Algeria

[1]http://orcid.org/ 0009-0002-7550-3369, [2]http://orcid.org/ 0000-0003-3108-6444, [3]http://orcid.org/ 0000-0002-4264-3376

Email: *toudert.bougrires@univ-biskra.dz, s.benmebarek@univ-biskra.dz, n.benmebarek@univ-biskra.dz.

ARTICLE INFO

Article History
Received: October 23, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Piled Raft Foundation,
Battered Piles,
Soil Confinement,
Lateral Resistance,
FLAC3D.

ABSTRACT

The piled raft (PR) foundation system combines a shallow raft with deep piles, mobilising both components through complex raft–pile–soil interactions. Due to its effectiveness under vertical and lateral loads, it is widely used to support tall, heavily loaded structures. However, the complexity of the system often leads to reliance on simplified design methods. This study uses numerical simulations in FLAC3D to investigate the effect of pile battering on the behaviour of PR and pile group (PG) foundations in drained sandy soils. Unlike previous research combining vertical and battered piles, only battered piles are adopted here to isolate their influence under combined loading. A batter angle of 10° is selected to avoid excessive bending moments. Results indicate that battered piles improve lateral resistance in PR foundations by up to 66%, though vertical capacity is slightly reduced (by 14%) due to axial–bending interaction. Under combined loading, vertical piles outperform battered ones in overall resistance gain (55.7% vs. 38.8%), mainly due to improved soil confinement. Batter piles also play a key role in reducing tilt and enhancing foundation stability. This study offers practical insights for designing foundations exposed to strong lateral forces, such as in offshore structures and wind turbines.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

List of notations

PR	Piled raft foundation	Lp	Lateral load carried by the piles in PR
BPR	Battered PR	Lr	Lateral load carried by the raft in PR
PG	Pile Group Foundation	γ'	Unit weight of the soil
BPG	Battered PG	φ	Friction angle of the soil
PRp	Piles component in PR	ψ	Dilation angle of the soil
PRr	Raft component in PR	δ	Soil/Foundation interface friction angle
Br	Width of raft	D_{50}	Mean grain size
Tr	Thickness of raft	E	Young's modulus
Lp	Length of pile	ν	Poisson's ratio
D	Pile diameter	c'	Cohesion
Θ	Pile inclination angle	*Kn*	Interface normal stiffness
Lpr	Lateral load carried by PR	*Ks*	Interface shear stiffness

I. INTRODUCTION

Piled raft foundations have gained significant attention in geotechnical engineering as an efficient solution that combines the advantages of both shallow and deep foundations. By integrating a raft with a group of piles, this system enhances load-bearing capacity while minimising differential settlement. Unlike conventional pile or raft foundations, the piled raft system utilises the combined contribution of both components—the raft provides additional bearing support, while the piles primarily control settlement—resulting in improved structural performance and cost-effectiveness [1]. The behaviour of a piled raft is governed by complex soil-structure interactions, including pile-soil, pile-pile, raft-soil, and pile-raft interactions [2-5]. Due to these interactions, accurately evaluating the bearing capacity of piled rafts is challenging and requires consideration of multiple design parameters and soil conditions. While numerous studies have focused on the performance of piled raft foundations under vertical loads [6-11]. Infrastructure such as high-rise buildings, bridges, wind turbines, and offshore platforms are also subjected to significant horizontal loads. Many researchers have stated that it is necessary to understand the response of piled rafts under lateral loading through both experimental investigations and numerical simulations to establish design criteria [12-15]. However, experimental studies on piled rafts subjected to combined vertical and horizontal loading remain limited.

Although recent research efforts have attempted to address this gap by analysing the combined effects of vertical and lateral loads on piled rafts [16], [17], a comprehensive design framework for such conditions has yet to be fully established. Recently, there has been growing interest in incorporating batter piles into piled raft foundations due to their promising potential in enhancing foundation stability. However, research on this topic remains scarce, with only a few experimental and numerical studies available. [18] conducted vertical and horizontal load tests on pile groups and piled rafts with and without batter piles in dry sand under 1g gravitational conditions. Their results demonstrated that batter piles improve resistance and reduce settlement compared to purely vertical piles. Finite element analyses corroborate these findings, offering more profound insight into the resistance mechanisms of these foundation systems. Similarly, [19] employed three-dimensional finite element models to investigate the seismic behaviour of pile groups with batter piles. Their study considered various pile configurations, including groups with a combination of vertical and inclined piles. Both hinged and fixed-headed pile conditions were analysed to assess the influence of pile-to-cap connections. Their results indicated that pile groups with batter piles exhibit greater horizontal stiffness than those composed solely of vertical piles. Other researchers have reached similar conclusions regarding the effectiveness of batter piles. [15], found that replacing vertical piles with batter piles reduces settlement and increases load-carrying capacity.

[20] experimentally investigated 3-pile and 6-pile batter pile foundations under vertical and lateral loads and found that batter piles enhance overall stiffness and reduce settlement. [21] analysed the effect of batter angles ranging from 10° to 30° under lateral loads, determining that an in-battered pile inclination of 15°–20° provides optimal resistance. [22] conducted dynamic load tests on 6-pile foundation models with and without batter piles, demonstrating that piled rafts incorporating batter piles are highly effective in minimising settlement and inclination under vertical dynamic loading. Although numerical simulations have proven to be a powerful tool for understanding foundation behaviour, relatively few numerical studies have explored the influence of batter piles. [23] developed a numerical analysis method to examine deformation and load distribution in piled raft foundations with batter piles. Their parametric studies confirmed that batter piles significantly enhance the deformation behaviour of piled rafts. Given this context, the present study aims to model and analyse the experimental work conducted by [24] while varying the batter pile for two foundation configurations: piled raft and pile group. The objective is to assess the impact of batter angle on vertical and lateral resistance, offering more details about their behaviour under combined loading. By considering the coupled action of vertical and horizontal loading, this study seeks to improve the understanding of failure mechanisms and key design parameters for piled raft foundations with batter piles. The analysis is performed using a three-dimensional numerical model, validated against Matsumoto's experimental study to ensure accuracy and reliability.

II. FINITE DIFFERENCE MODELING

II.1 DESCRIPTION OF THE ANALYSED FOUNDATION SYSTEM

This study investigates the influence of pile inclination on the behaviour of piled raft and pile group foundations subjected to vertical, lateral, and combined loading in sandy soil. The problem is addressed numerically using a three-dimensional finite-difference model, which explicitly captures the interaction between the raft, the inclined piles, and the surrounding soil continuum. Special emphasis is placed on how pile inclination modifies load transfer mechanisms, stiffness, and failure patterns compared with conventional systems supported only by vertical piles. To isolate the mechanical response of inclined piles under combined loading, the foundation system is idealised as consisting solely of inclined piles rather than a mixed arrangement of vertical and inclined elements. This simplification allows a clearer interpretation of the role of inclination angle on global response, raft displacements, and pile forces, without the additional complexity introduced by heterogeneous pile layouts. The numerical framework enables systematic variation of inclination angle and load combination, thereby providing a consistent basis for parametric analysis.

The influence of pile inclination on both vertical and lateral load-carrying capacities is examined using displacement-based failure criteria. For vertical capacity, ultimate bearing capacity is often defined at a settlement equal to 10% of the raft width Br; however, this criterion is not directly suitable for the reduced-scale geometry adopted here. In line with recommendations for small-scale modelling, a stricter settlement limit equal to 10% of the pile diameter d is adopted to define the ultimate vertical capacity, ensuring that the response remains within an appropriate deformation range for model-scale interpretation. For lateral capacity Lpr, a horizontal displacement criterion of 0.1d is employed, consistent with common practice in evaluating laterally loaded pile and piled-raft systems. Within this framework, the effects of inclination are quantified in terms of: (i) ultimate vertical and lateral capacities, (ii) load–displacement behaviour under pure and combined loading, and (iii) load sharing between raft and piles as the inclination angle and loading regime vary. The different external load components and the internal forces mobilised in the piled-raft system under combined vertical–lateral loading are schematically illustrated in Figure 1, providing a conceptual basis for the subsequent numerical analyses.

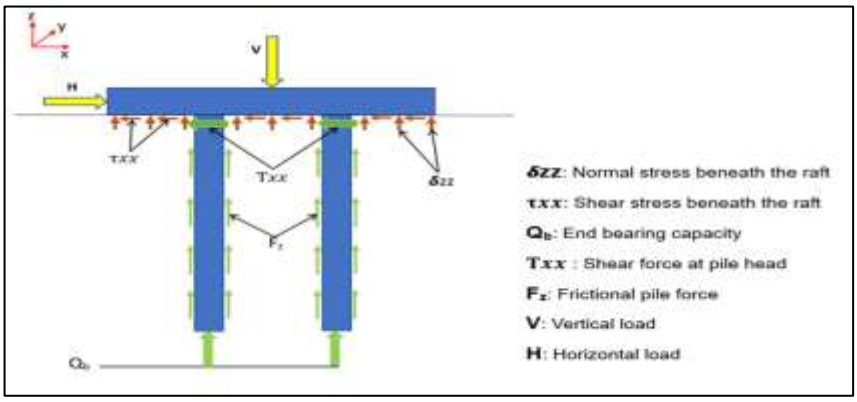


Figure 1: Forces Developed at the Soil-Piled Raft Interface.
Source: Authors, (2026).

The lateral load contribution of the raft (Lr) in the piled raft system is determined by integrating the shear stress at the interface between the soil and the raft. Figure 1 illustrates this, and it is expressed in Equation (1). In these equations, τiixx represents the shear stress at the interface between the soil and the raft for each element i, and dSi is the segment's cross-sectional area of element i.

$$Lr = \iint_{-Br/2}^{Br/2} \tau ixx \ \ dSi \tag{1}$$

Due to the complexity of directly calculating the lateral load contribution of piles, especially inclined piles, it is determined indirectly by deduction. The lateral load carried by the piles (Lp) is calculated as the total lateral load of the piled raft system (Lpr) minus the lateral contribution of the raft (Lr), as expressed in Equation (2):

$$Lp = Lpr - Lr \tag{2}$$

Two foundation configurations are considered: a piled raft foundation where the raft is in direct contact with the soil, and a pile group foundation where the raft is elevated above the ground by a distance equal to the raft thickness. This distinction allows for a clear comparison between configurations with and without raft–soil contact, emphasising the role of the raft in load sharing and settlement control. The effect of pile inclination is examined by comparing vertical (0°) and inclined (10°) pile arrangements. Four configurations were studied, as shown in **Figure 2**, to analyse the combined effect of pile inclination and foundation type on load distribution, settlement behaviour, and lateral resistance. By evaluating the performance of these foundation systems, this study aims to provide insights into the role of pile inclination in optimising foundation design, particularly in sandy soils where both vertical and lateral loads play a crucial role in structural stability.

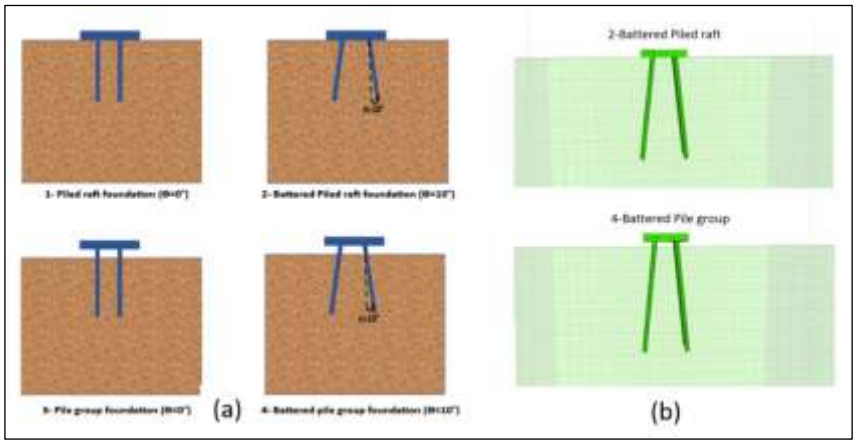


Figure 2: (a) Configurations studied in this paper, (b) battered confugurations modeled in FLAC 3D.
Source: Authors, (2026).

II.2 MODELING PROCEDURE

This paper conducts numerical simulations using the explicit finite difference code FLAC3D (Itasca Consulting Group Inc., 2013) to assess the behaviour of a battered piled raft under combined vertical and lateral loading in sandy soil. FLAC3D offers robust computational capabilities, making it particularly valuable for analysing foundation systems, where experimental testing, especially on inclined piles, is complex and resource-intensive. Additionally, 1g experimental tests are known to suffer from scale effects and limitations in reproducing real stress conditions, justifying the use of numerical modelling to achieve accurate and scalable results. The numerical model is based on the experimental work of [24] which involved 1g tests on composite piled raft foundations subjected to vertical and lateral loads. The experimental setup consisted of a 400 mm × 400 mm raft, 40 mm thick, supported by four piles of 600 mm length and 40 mm diameter, arranged with uniform spacing of 200 mm in both the x and y directions. The piles were embedded in a two-layer soil profile consisting of 1 metre of Toyoura sand overlying a 1-metre-thick compacted brick base.

While the primary goal is to simulate [24] experimental configuration for validation, the study also aims to contribute to the broader engineering objective of optimising foundation systems for high-rise buildings and structures subject to combined vertical and lateral loads—such as wind turbines and tall towers—by evaluating the effect of pile inclination on load capacity and load sharing. In the simulations, the piles are modelled using six-node cylindrical solid elements, while the raft is represented by a combination of 12-node radial cylindrical solid elements and eight-node brick elements. Although solid elements do not directly capture internal forces, they allow a realistic geometric representation of structural components under external loads. The overall soil domain dimensions are set to 3 m × 3 m × 2 m—twice the lateral extent of the experimental setup—to reduce the influence of boundary effects. Boundary conditions include fixed constraints at the base and lateral restrictions against displacement in the X and Y directions.

The soil is modelled using the Mohr–Coulomb failure criterion, selected for its simplicity and effectiveness in capturing the essential behaviour of sandy soils with minimal input parameters. Although more advanced constitutive models such as the Hardening Soil model exist, the Mohr–Coulomb model is adequate for this study's scope and offers a balance between computational efficiency and realistic approximation of soil strength. The parameters used to simulate Toyoura sand are Young's modulus E = 44 MPa, Poisson's ratio $\nu = 0.3$, unit weight $\gamma' = 15.9$ kN/m3, friction angle $\phi' = 40°$, dilation angle $\psi = 10°$ and cohesion $c' = 0$ kPa. The parameters for the soil and foundation are summarised in Table 1.

Table 1: Material parameters used in [24] and this study.

Parameter	Toyoura sand	Base brick	Foundation
E (MPa)	44	$6x10^3$	$7x10^4$
ν	0,3	0,2	0.3
γ' (kN/m3)	15,90	22	27
c'(kPa)	0	/	/
φ' (°)	40	/	/
ψ (°)	10	/	/
D 50 (mm)	0,17	/	/

Source: [24].

The mesh was carefully refined beneath and around the raft and piles using a fine grid with an element size approaching the average particle diameter of Toyoura sand ($D_{50} = 0.17$ mm). A mesh convergence study was performed and showed that stable, mesh-independent results were obtained when the minimum element size was taken approximately equal to D_{50}, which was therefore adopted in the final model. Mesh refinement was continued until changes in key response quantities (such as raft displacement and pile head shear) between successive refinements fell within a small tolerance, ensuring that the numerical response was controlled by the physics rather than by the discretization. The mesh gradually becomes coarser with distance from the foundation to reduce computational cost while maintaining numerical accuracy. The raft and piles are assumed to be isotropic and perfectly elastic. Soil–structure interaction is simulated using interface elements that allow for realistic relative movement between soil and structural components. Interface properties were defined as follows: friction angle $\delta = 40°$, cohesion $c = 0$ kPa, normal stiffness $Kn = 10^9$ Pa/m, and shear stiffness $Ks = 10^9$ Pa/m.

These interface stiffness values are selected to be sufficiently high to ensure force transfer without inducing unrealistic rigidity, and sensitivity checks confirmed that moderate variations in Kn and Ks had negligible influence on the global response. The loading process was divided into four main steps: (1) gravity loading to generate the initial geostatic stress field; (2) installation of the raft and piles by assigning structural material properties to the relevant zones; (3) application of a uniform vertical load on the raft to simulate a surcharge; and (4) application of a lateral force at the centre of the raft's vertical surface. Vertical and lateral loads were applied incrementally, with the vertical load kept constant during the lateral loading phase to represent combined loading and to capture the progressive development of nonlinearity in the soil response. The numerical model mesh and the foundation dimensions are illustrated in Figure 3.

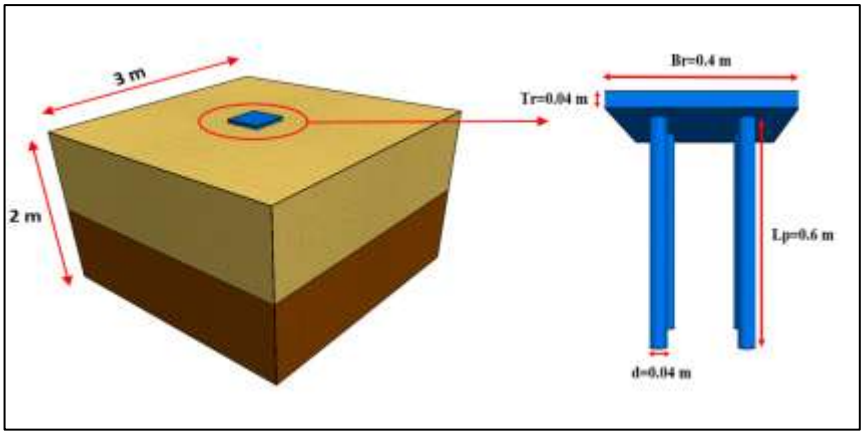


Figure 3: Mesh Pattern and Element Dimensions.
Source: Authors, (2026).

II.3 VALIDATION OF NUMERICAL MODELING

To validate the numerical modelling procedure used in this study, the results were compared with one case study from the seven experimental tests conducted by [24]. In this case, a rigid piled raft foundation was tested under static vertical and static horizontal loading. The loading process was divided into two stages: the vertical loading stage and the cyclic horizontal loading stage. In the vertical loading stage, a total vertical load of Vi =3.384 kN was applied to the model foundation. After completing the vertical stage, a cyclic horizontal loading stage was applied to the model raft. It is important to note that, although cyclic horizontal loading was performed on the model foundation, a monotonic horizontal loading test was also carried out. As illustrated in Figure 4, the numerical results closely align with the experimental data, demonstrating the accuracy of the simulations.

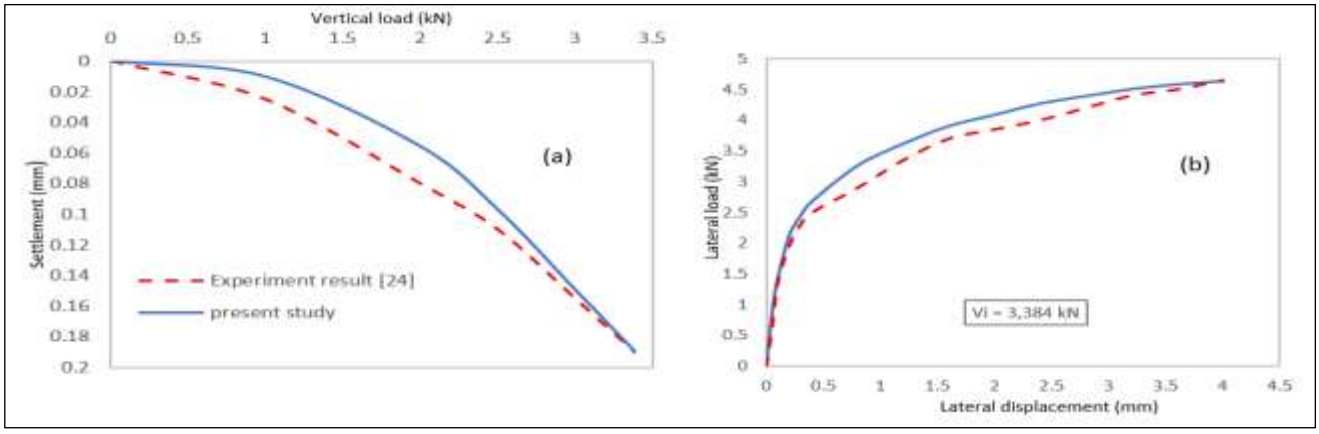


Figure 4: Comparison between the results of present study and previous experimental work [24], (a) load-settlements and (b) lateral load-lateral displacement behaviors of piled raft.
Source: [24].

III. RESULTS AND DISCUSSIONS

III.1 PURE VERTICAL LOAD TEST

The influence of pile inclination on vertical resistance was first examined. As shown in Figure 5, increasing the inclination angle from 0° to 10° negatively impacts vertical resistance for both foundation types. Under a vertical load of 3 kN, settlement increases from 0.14 mm to 0.16 mm for the piled raft foundation and from 0.27 mm to 0.316 mm for the pile group foundation. This reduction in vertical resistance occurs because a portion of the applied vertical load is redirected into a horizontal component, reducing the effective axial capacity of the piles. As a result, settlement increases. In piled raft foundations, the decreased pile contribution transfers more load onto the raft, leading to greater deformation. In pile group foundations, inclined piles experience additional bending stresses, further diminishing their ability to resist vertical loads. Consequently, as the inclination angle increases, vertical resistance decreases.

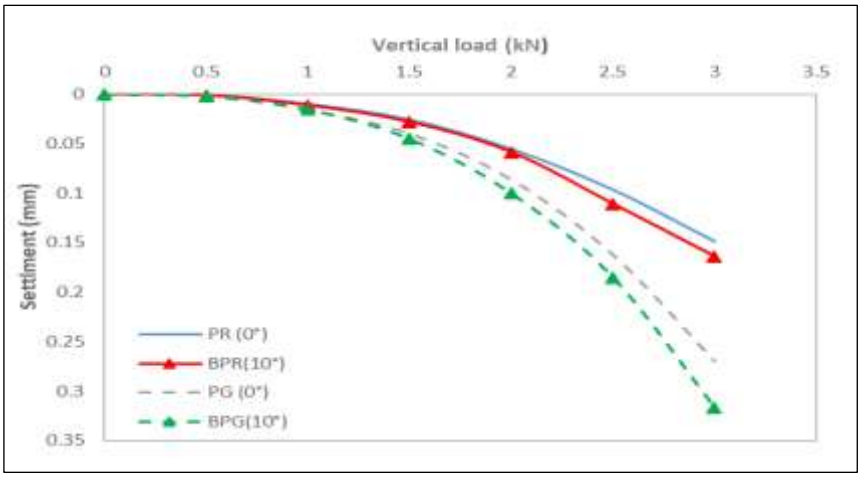


Figure 5: Vertical Load vs settlement curves of PR, BPR.PG and BPG.
Source: Authors, (2026).

III.2 PURE LATERAL LOAD TEST

In this study, the foundation was initially subjected only to the force of gravity due to its own weight, without any external loads applied. We then analysed the effect of pile inclination on lateral resistance by applying a pure lateral load until a lateral displacement equal to 10% of the pile's diameter was reached. This specific load, as corroborated by several studies, represents the lateral resistance of the foundation. As illustrated in Figure 6, under pure lateral loading, lateral resistance consistently increases with the inclination angle. For the piled raft foundation, increasing the inclination angle from 0° to 10° led to a significant 66% increase in lateral resistance, rising from 2.853 kN to 4.74 kN. In contrast, the increase for the pile group foundation was more modest, with lateral resistance increasing by

21%, from 2.864 kN to 3.478 kN. The increase in lateral resistance with pile inclination is due to the enhanced pile-raft-soil interaction. When piles are inclined, they interact differently with the surrounding soil compared to vertical piles. Instead of resisting the lateral load purely through bending, inclined piles transfer part of this load into axial compression or tension. This redistribution reduces bending stress and increases overall stability. Additionally, a portion of the lateral load is redistributed into axial compression or tension along the pile length, reducing bending effects and enhancing stiffness. This inclination also modifies the pile-soil contact, activating greater lateral resistance, particularly in sandy soil. In a piled raft foundation, the raft's contribution becomes more significant as the axial load component transferred from the inclined piles increases the raft-soil interaction. This combined effect of pile-raft-soil interaction and load redistribution explains the improved lateral resistance observed with inclined piles.

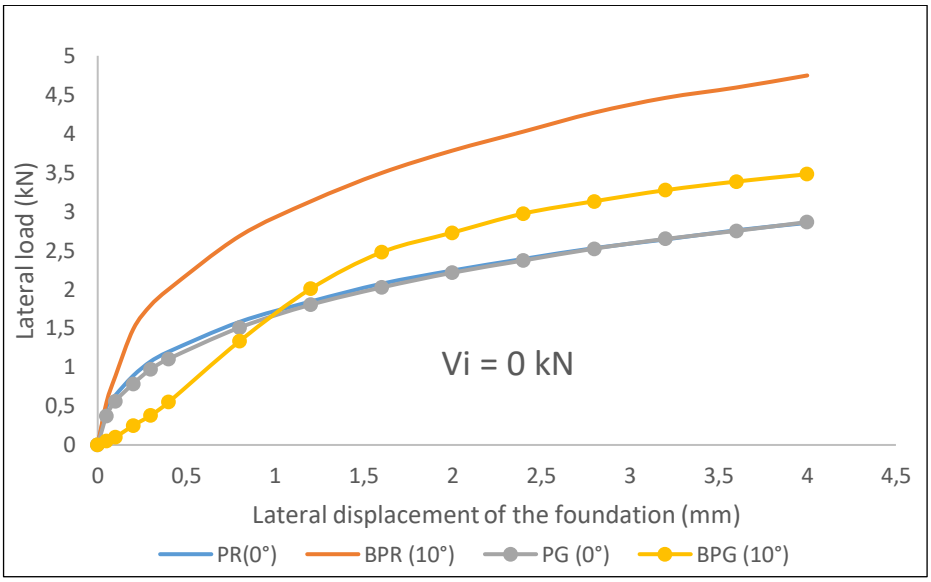


Figure 6: Lateral load vs lateral displacement under pure lateral load curves of PR, BPR.PG and BPG.
Source: Authors, (2026).

III.3 COMBINED VERTICAL AND LATERAL LOADS TEST

Finally, the behaviour of both foundation types was analysed under a combined loading. For the four configurations mentioned earlier, a vertical load of Vi= 3 kN was first applied until stabilisation, followed by a lateral load until a lateral displacement equal to 10% of the pile diameter was reached. As shown in Figure 6 and Figure 7, lateral resistance under combined loading is higher than under pure lateral loading. For vertical piles in the piled raft foundation, lateral resistance reaches 4.44 kN under combined loading, representing a 55.7% increase compared to pure lateral loading. For inclined piles, lateral resistance under combined loading reaches 6.59 kN, whereas it was 4.74 kN under pure lateral loading, indicating a 38.8% increase. Similarly, for the pile group foundation, lateral resistance increases by 16.23% and 27.4% when the inclination angle changes from 0° to 10°, respectively. This difference in percentage increase can be attributed to the role of vertical load in enhancing soil confinement and increasing normal stress along the pile shaft, which is more beneficial for vertical piles. Additionally, vertical piles rely primarily on passive soil resistance and shaft friction for lateral resistance, while inclined piles already experience higher lateral stiffness, making the effect of vertical load less pronounced.

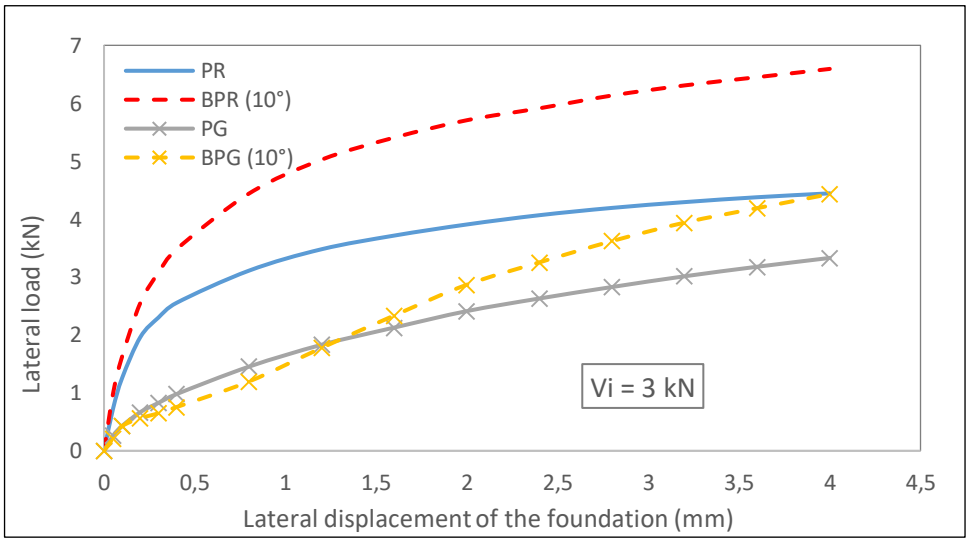


Figure 7: Lateral load vs lateral displacement under combined loads (Vi=3N) of PR, BPR.PG and BPG.
Source: Authors, (2026).

III.4 LOAD DISTRIBUTION AND LATERAL RESISTANCE UNDER PURE AND COMBINED LOADINGS

III.4.1 Under Pure Lateral Loading

Figure 8 shows that the load-sharing mechanism between the raft and piles under pure lateral load remains constant regardless of pile inclination; however, the contribution of each component varies. As pile inclination increases, both the raft and piles mobilize increased lateral resistance. The lateral resistance of the raft increases by 48% due to greater soil-raft interaction, with inclined piles transferring more load to the raft. Meanwhile, the 69% increase in pile resistance suggests that inclined piles engage a greater soil mass, thus improving their ability to resist lateral loads. It is important to highlight the complexity of the soil-raft interaction. Comparing piles in the PR system with piles in the PG system, when the piles in the PR system are vertical, they support 89% of the load carried by the piles in the PG system. However, at an inclination of 10°, this percentage reaches 123%. This increase results from the higher axial forces developed by inclined piles due to their orientation, which improves the efficiency of load transfer. Therefore, inclined piles not only improve lateral resistance, but also significantly change the overall load distribution within the foundation system.

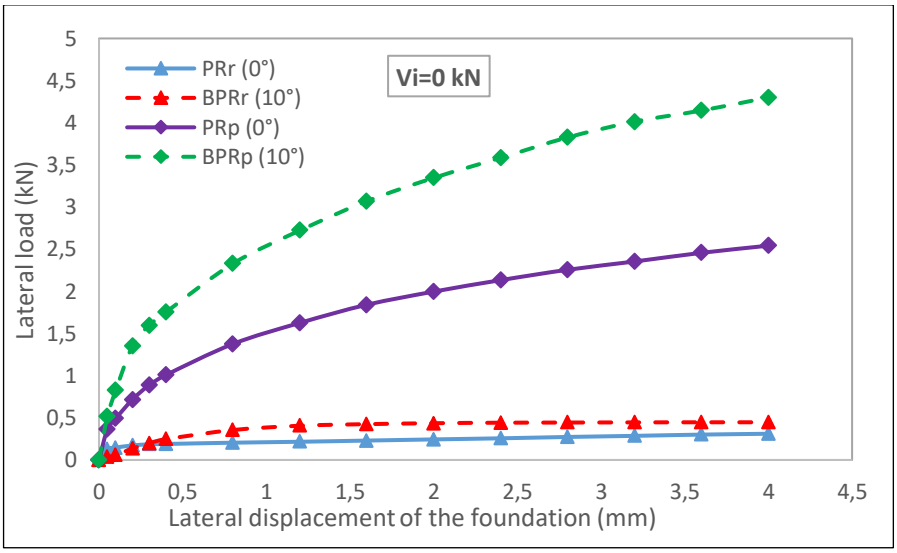


Figure 8: Distribution of Lateral load vs lateral displacement curves for PRr, BPRr, PRp and BPRp under pure lateral load.
Source: Authors, (2026).

III.4.2 Under Combined Vertical and Lateral Loads

Under combined loading, increasing pile inclination enhances lateral resistance for both foundation types. As shown in Figure 9, the lateral contribution of the raft increases slightly from 0.641 kN to 0.691 kN, while the pile contribution rises more significantly from 3.8 kN to 5.9 kN. While comparing the lateral resistance of vertical and inclined piles in PR against that of PG. For vertical piles, the lateral resistance in the PR is 3.8 kN, slightly higher than the 3.32 kN observed for PG. Similarly, for inclined piles, the piles in BPR exhibit greater lateral resistance 5.9 kN compared to BPG 4.4 kN. These results indicate that under vertical loading, both vertical and inclined piles in a PR exhibit improved lateral performance due to enhanced interaction with the surrounding soil.

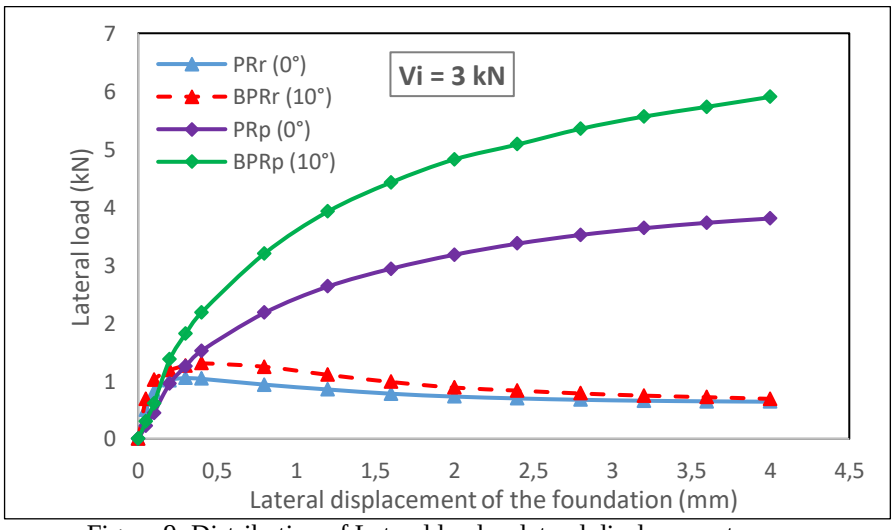


Figure 9: Distribution of Lateral load vs lateral displacement curves under combined loads (Vi=3N) for PRr, BPRr, PRp and BPRp.
Source: Authors, (2026).

III.5 RAFT INCLINATION BEHAVIOR UNDER LATERAL AND COMBINED LOADS

As shown in Figure 10, when the foundation is subjected to pure lateral load, the raft in the piled raft system tilts slightly less than in the pile group, with inclinations of 7.824 × 10^{-3} radians and 8.099 × 10^{-3} radians, respectively. This suggests that the piled raft performs a bit better in controlling raft inclination under lateral forces. When inclined (or battered) piles are used, the raft inclination decreases even further. In the case of the battered piled raft, the inclination drops to 6.749 × 10^{-3} radians, and for the battered pile group, it goes down to 7.46 × 10^{-3} radians. This shows how effective inclined piles can be in improving the stability and stiffness of the foundation under lateral loading. When both vertical and lateral loads are applied (combined loading), the vertical load plays an important role in further reducing the raft's inclination. We can see this in all four foundation configurations: the piled raft shows an inclination of 7.274 × 10^{-3} radians, the battered piled raft 6.749 × 10^{-3} radians, the pile group 7.499 × 10^{-3} radians, and the battered pile group 6.824 × 10^{-3} radians. Overall, these results highlight how vertical load helps stabilise the foundation, and they confirm that the battered piled raft performs the best at minimising raft inclination under combined loading conditions.

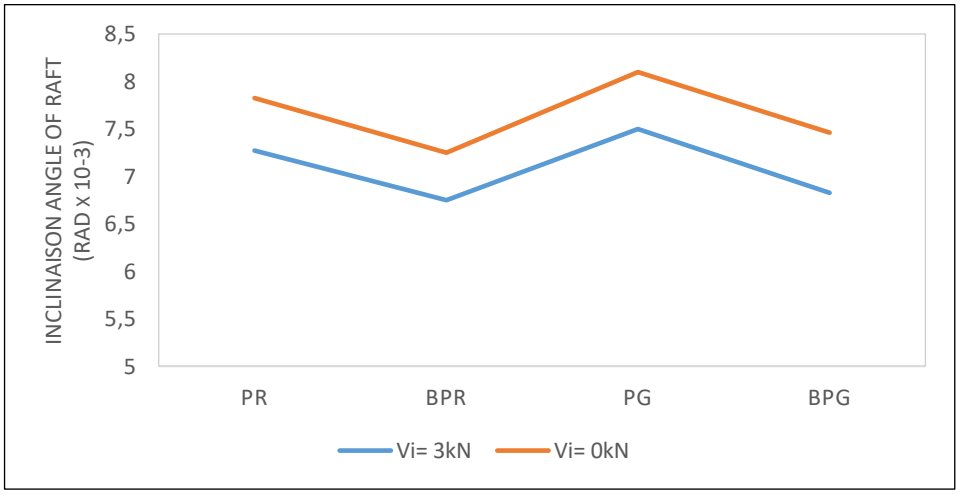


Figure 10: Inclination angles of raft under pure lateral and combined loads for PR, BPR, PG and BPG.
Source: Authors, (2026).

IV. CONCLUSIONS

This study concludes that pile inclination has a dual impact on piled raft and pile group foundations: it reduces vertical load capacity and increases settlements due to bending and axial load redistribution, especially under pure vertical loading, but markedly enhances lateral resistance under pure lateral and combined loading by mobilizing larger soil volumes and redistributing forces more effectively. The combined load scenario reveals that vertical piles benefit more from vertical load-induced soil confinement than inclined piles, though both configurations improve lateral stiffness. Load sharing shifts with inclination, where inclined piles significantly increase raft and pile resistance and carry a higher portion of total load. Raft inclination under lateral and combined loading decreases most with battered piles, highlighting their stabilizing effect. Overall, while inclined piles compromise vertical capacity, their significant lateral resistance gains make them critical for foundations subject to lateral or combined loads, affirming pile inclination as a key design parameter consistent with the research objectives and scientific considerations outlined in this work.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Toudert Bougrires and Sadok Benmebarek.
Methodology: Toudert Bougrires and Sadok Benmebarek.
Investigation: Toudert Bougrires.
Discussion of results: Toudert Bougrires, Sadok Benmebarek, and Naïma Benmebarek.
Writing – Original Draft: Toudert Bougrires.
Writing – Review and Editing: Toudert Bougrires, Sadok Benmebarek and Naïma Benmebarek.
Resources: Toudert Bougrires, Sadok Benmebarek.
Supervision: Sadok Benmebarek and Naïma Benmebarek.
Approval of the final text: Toudert Bougrires, Sadok Benmebarek, and Naïma Benmebarek**.**

VI. ACKNOWLEDGMENTS

The authors express their gratitude to the Directorate General for Scientific Research and Technological Development of Algeria.

VII. REFERENCES

[1] H. G. Poulos, "Piled raft foundations: design and applications," Geotechnique, vol. 51, no. 2, pp. 95–113, 2001.

[2] R. Katzenbach, U. Arslan, and C. Moormann, "Piled raft foundation projects in Germany," in Design Applications of Raft Foundations, Thomas Telford Publishing, pp. 323–391, 2000.

[3] L. de Sanctis and G. Russo, “Analysis and performance of piled rafts designed using innovative criteria,” J. Geotech. Geoenviron. Eng., vol. 134, no. 8, pp. 1118–1128, 2008.

[4] R. Katzenbach, D. Choudhury, and D. W. Chang, “General report of TC212 deep foundations,” in Proc. 18th Int. Conf. Soil Mech. Geotech. Eng., pp. 2651–2658, 2013.

[5] A. Kumar, D. Choudhury, and R. Katzenbach, “Behaviour of combined pile-raft foundation (CPRF) under static and pseudo-static conditions using PLAXIS3D,” in Proc. 6th Int. Conf. Earthquake Geotech. Eng. (6ICEGE), Christchurch, New Zealand, Paper ID-140, 2015.

[6] L. de Sanctis and A. Mandolini, “Bearing capacity of piled rafts on soft clay soils,” J. Geotech. Geoenviron. Eng., vol. 132, no. 12, pp. 1600–1610, 2006.

[7] R. Banerjee, S. Bandyopadhyay, A. Sengupta, and G. R. Reddy, “Settlement behaviour of a pile raft subjected to vertical loadings in multilayered soil,” Geomech. Geoeng., vol. 17, no. 1, pp. 282–296, 2022.

[8] R. S. Al Masoodi, A. M. Al Gharrawi, and H. H. Hussein, “The behaviour of piled raft foundation in loose sandy soil,” J. Phys.: Conf. Ser., vol. 1973, no. 1, p. 012001, 2021.

[9] A. Ferchat, S. Benmebarek, and M. N. Houhou, “3D numerical analysis of piled raft interaction in drained soft clay conditions,” Arab. J. Geosci., vol. 14, pp. 1–12, 2021.

[10] R. Modak and B. Singh, “A parametric study of large piled raft foundations on clay soil,” Ocean Eng., vol. 262, p. 112251, 2022.

[11] N. Bralović, I. Despotović, and D. Kukaras, “Experimental analysis of the behaviour of piled raft foundations in loose sand,” Appl. Sci., vol. 13, no. 1, p. 546, 2022.

[12] R. Katzenbach and J. Turek, “Combined pile-raft foundation subjected to lateral loads,” in Proc. 16th Int. Conf. Soil Mech. Geotech. Eng., IOS Press, pp. 2001–2004, 2005.

[13] D. Chanda, R. Saha, and S. Haldar, “Behaviour of piled raft foundation in sand subjected to combined VMH loading,” Ocean Eng., vol. 216, p. 107596, 2020.

[14] U. Kumar and S. Vasanwala, “A numerical analysis on the effect of pile head connections on piled raft foundation subjected to vertical and static horizontal load,” Mater. Today: Proc., vol. 42, pp. 3083–3088, 2021.

[15] I. Jamil, I. Ahmad, I. Khan, W. Ullah, A. U. Rehman, and S. A. Khan, “Factors affecting the lateral contribution of a raft in a piled raft system,” Ain Shams Eng. J., vol. 14, no. 5, p. 101968, 2023.

[16] Y. S. Unsever, T. Matsumoto, and M. Y. Özkan, “Numerical analyses of load tests on model foundations in dry sand,” Comput. Geotech., vol. 63, pp. 255–266, 2015.

[17] P. Deb and S. K. Pal, “Nonlinear analysis of lateral load sharing response of piled raft subjected to combined VL loading,” Mar. Georesour. Geotech., vol. 39, no. 8, pp. 994–1014, 2021.

[18] A. T. Vu, T. Matsumoto, S. I. Kobayashi, and T. L. Nguyen, “Model load tests on battered pile foundations and finite-element analysis,” Int. J. Phys. Model. Geotech., vol. 18, no. 1, pp. 33–54, 2018.

[19] A. Giannakou, N. Gerolymos, G. Gazetas, T. Tazoh, and I. Anastasopoulos, “Seismic behavior of batter piles: elastic response,” J. Geotech. Geoenviron. Eng., vol. 136, no. 9, pp. 1187–1199, 2010.

[20] A. T. Vu, T. Matsumoto, S. I. Kobayashi, and S. Shimono, “Experimental study on pile foundations having batter piles subjected to combination of vertical and horizontal loading at 1-g field,” Geotech. Eng. J. SEAGS & AGSSEA, vol. 48, no. 3, pp. 12–24, 2017.

[21] G. Mahalakshmi and B. Soundara, “Numerical analysis of batter piles under lateral loading,” Int. J. Geotech. Eng., vol. 15, no. 5, pp. 555–563, 2021.

[22] A. T. Vu and T. Matsumoto, “Dynamic load tests on piled raft foundations with batter piles,” Geotech. Eng. J. SEAGS & AGSSEA, vol. 51, no. 1, pp. 1–10, 2020.

[23] P. Kitiyodom and T. Matsumoto, “A numerical study on the behavior of piled raft foundations with batter piles,” Soils Found., vol. 42, no. 4, pp. 1–13, 2002.

[24] T. Matsumoto et al., “Load tests of piled raft models with different pile head connection conditions and their analyses,” Soils Found., vol. 50, no. 1, pp. 63–81, 2010.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 176-187. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2851

RESEARCH ARTICLE OPEN ACCESS

INFLUENCE OF DEFECTS ON THE MECHANICAL PROPERTIES OF METALLIC COATINGS OBTAINED BY THERMAL SPRAYING: A REVIEW

Uirá Rodrigues da Silveira Montani[1], Marília Garcia Diniz[2]
Paolo Bettini[3] and Hector Reynaldo Meneses Costa[4]

[1]Programa de Pós-Graduação em Engenharia Mecânica e dos Materiais do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (PPEMM - CEFET-RJ), Rio de Janeiro, Brazil
[2]Programa de Pós-Graduação em Engenharia Mecânica da Universidade do Estado do Rio de Janeiro (PPGEM – UERJ), Rio de Janeiro, Brazil
[3, 4]Aerospace Science and Technology Department of Politecnico Di Milano, Milan, Italy.

[1]https://orcid.org/0009-0005-8174-9164, [2]https://orcid.org/0000-0002-1836-8618
[3]https://orcid.org/0000-0003-1083-5792, [4]https://orcid.org/0000-0002-0887-6838

Email: uirarsilveira@yahoo.com.br, mgarciadiniz@gmail.com, paolo.bettini@polimi.it, hectorey@gmail.com

ARTICLE INFO

Article History
Received: October 24, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Coatings,
Thermal Spray,
Defects,
Machine Learning

ABSTRACT

Thermally sprayed coatings are used as superior engineering materials, to prolong the service life of components and reduce manufacturing costs, but also significantly saving limited resources and energy, a technique capable to protect surfaces from erosive wear, corrosion, and enhance the tribological properties of a broad range of engineering materials. The study outlines the technological principle and process of thermal spray technique, and explores microstructural defects of thermal sprayed coatings, including pores, microcracks, unmelted particles, and oxides, exploring their formation mechanisms and effects. This paper also explores machine learning as a valuable tool that enables more accurate prediction of defect formation and supports the development of optimized spraying parameters.

Copyright ©2025 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Thermal spray coatings are widely applied in engines, pumps, compressors, pipelines, and medical implants, providing cost-effective and reliable means to protect, repair, and extend component lifespan. However, inherent defects arise from compositional imbalances and surface irregularities generated during the layer-by-layer deposition process, leading to stress concentration and crack formation. These defects, influenced by the feedstock material and deposition parameters, reduce coating density and compactness, thereby impairing oxidation resistance and overall performance. This paper aims to present a comprehensive review of the current state of thermal spraying, beginning with an overview of existing technologies and recent advancements, followed by a detailed analysis of the main defects observed in such coatings. It also discusses current challenges in the field and identifies key research gaps, contributing to a deeper understanding and guiding future investigations in thermal spray technology.

II. METHODOLOGY

The literature survey was conducted as part of the development of this research paper. This study examined different thermal spray processes and the formation of defects on the surface and within the core of metallic and ceramic coatings. The literature review provides a strong foundation for the current research and highlights its relevance for future investigations. The study relied on references from various national and international journals, articles, and reputable websites.

III. LITERATURE REVIEW

III.1 THERMAL SPRAY

The thermal spray process (TS) involves several distinct deposition techniques utilizing a concentrated heat source to liquefy different material compounds. These molten particles are then propelled onto the surface to be coated, each carrying different levels of kinetic energy [1]. The particles coalesce into microstructural artifacts resembling splats, which solidify into an interconnected network of a lamellar microstructure, typically resulting in a functional coating ranging from 50 µm to several millimeters in thickness [1-5] (see Figure 1). The thermal spray technique is a surface modification process that enhances the substrate material characteristics, cappable of supplying the underlying substrate with good thermal insulation, thermally stable, wear resistant, high temperature oxidation resistant and hard coating [2-4]. Rigamonti et al. [5] proposed a method for embedding polymer-coated optical fibers with inscribed fiber Bragg gratings (FBGs) onto steel substrates using the thermal spray technique. This approach ensured strong adhesion, efficient strain transfer, and minimal invasiveness, while preserving the integrity of the fiber's core and cladding. The thermal spray process, performed at low temperatures, enabled the integration of the fiber into the metallic structure without causing damage, and the resulting metal layer provided additional mechanical protection, allowing the system to be used in harsh environments characterized by severe corrosion and wear.

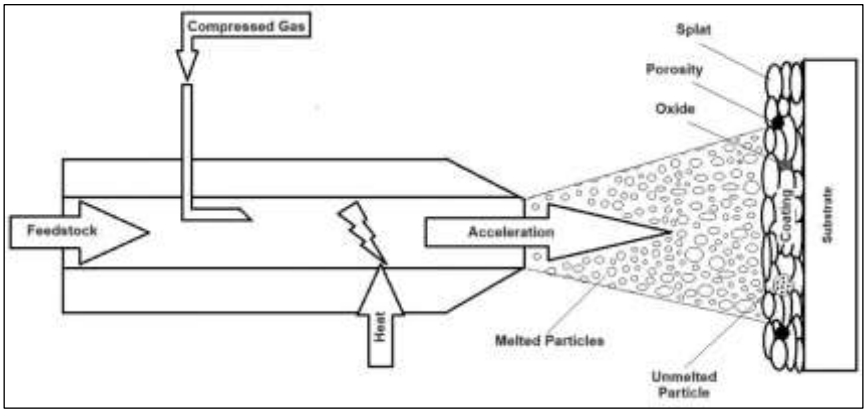


Figure 1: Thermal Spray Process.
Source: Authors, (2026).

Thermal spray processes are typically grouped according to their energy source, as illustrated in Figure. 2. These processes are differentiated based on the temperature and velocity of the jet stream they generate. Splats are frequently observed in thermally sprayed coatings, with the characteristics of splat boundaries (often oxides at these interfaces), porosity, and residual stress varying significantly. Moreover, corrosive substances can exploit the interplat bonding and interconnected pores as pathways for short-circuit diffusion [4].

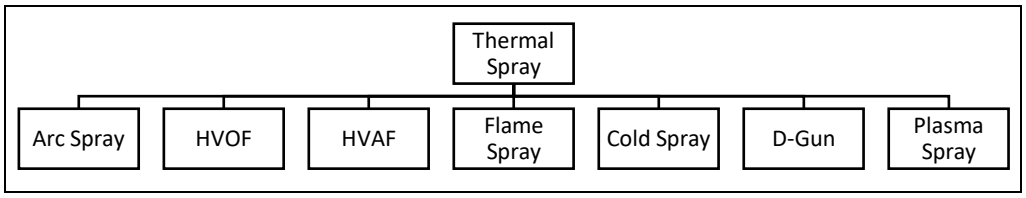


Figure 2: Generalized classification of thermal spray techniques.
Source: Authors, (2026).

Surface modifications using coating technologies serve to safeguard the substrate material, diminishing replacement expenses by slowing down material degradation, consequently prolonging the lifespan of components [1],[2]. Coating methods are categorized based on specific needs, including desired coating thickness, adhesion mechanism, necessary mechanical strength, component geometry and chemistry, coating process conditions suitable for the component, and the operating conditions of the coating [2]. Thermal spray processes have been extensively validated across industries such as steel, power generation, automotive, aerospace, marine, and petrochemical sectors for safeguarding and repairing components [4-7]. Utilizing thermal spray for repairing and remanufacturing components with localized defects or failures not only extends their service life but also proves to be cost-effective. This approach significantly conserves resources and energy, aligning well with the core principles of sustainable development [6]. The metallic coatings manufactured through additive manufacturing (AM) processes, such as thermal spray, often exhibit irregular surface morphology compared to those produced by conventional methods.

These irregularities stem from the layer-by-layer deposition and fusion occurring on the material surfaces [8], [9]. Several factors contribute to the irregularity of these surfaces, including the staircase effect, agglomeration of partially fused material, spattering, splashed particles (evaporation and balling effect), and the instability of the molten pool (wetting effect). These features formed on the surfaces act as stress concentration sites, facilitating crack formation. Furthermore, the size of the feedstock and deposition parameters also influence the surface morphology of the coatings [9],[10]. Typically, metallic alloys are not thermodynamically stable; compositional imbalances arise due to metallurgical transformations occurring in the structures, retaining crystallographic imperfections or defects [8],[11]. The commonly found defects in single-material-printed parts are microcracks, porosity, unmelted or solidified particles. Bonding defects arise from inadequate energy input, potentially leading to pore formation at the interface [8]. Cavities and micro-cracks stem from shrinkage, whereas pores result from anomalous gas entrapment within the molten material. In cold spray additive manufacturing (AM)

processes, such as deposition, insufficient bonding between particles and inefficient plastic deformation may give rise to the formation of micro-pores and inter-particle boundaries at lower impact velocities [9]. The objective of this paper is to comprehensively review the current state of knowledge in thermal spraying. It begins by providing an overview of thermal spraying technologies, highlighting recent advancements in the field. Subsequently, it conducts an in-depth analysis of various defects encountered in thermal spray coatings. Finally, the paper addresses the current challenges and issues encountered by researchers in this domain. The authors anticipate that this manuscript will contribute significantly to the existing body of knowledge by offering a detailed examination and identification of challenges for future research endeavors.

III.2 DEFECTS IN COATINGS

III.2.1 Porosity

Porosity refers to the volume of pores relative to that of the bulk material, and its increase indicates a rise in the volume of pores within the entire structure [12]. Typically, porosity ranges between 0.1 and 15% for thermal spray coatings [13]. It is an inherent characteristic of coatings. Numerous factors contribute to the porosity features in coatings during the preparation process of thermal-spraying coatings, including the impact energy of incompletely melted particles, the solidification and shrinkage of the coating, the shielding effect caused by different spray angles, and the wide gap in droplet velocity. Additionally, entrapped gas inclusions from powder and temperature play roles in porosity characteristics in coatings, such as morphology, pore diameters, pore distributions, microcrack sizes and orientations, and lamellar splats [13],[14]. Furthermore, the types of components and bonding mechanisms impact pore distributions [4].

The temperature of incoming droplets can vary from fully molten liquid to a completely solid state, depending on the particle's temperature. Liquid droplets flow freely and cover most of the spaces. The impact velocity of solid or partially melted particles must be sufficiently high to deform plastically upon impact. Consequently, particles with higher velocities undergo greater deformation, resulting in finer void closure (as seen in Figure 3). If the impact velocity is modest, solid particles may become lodged in the coating due to subsequent particles approaching later. These unformed particles have poor adhesion and should be further apart from the underlying splat, forming voids [4],[13]. Porosity reflects the fundamental properties of the coating, and typically, lower porosity indicates better coating formation and quality properties. Excessive porosity can reduce strength, abrasion resistance, and corrosion resistance [13].

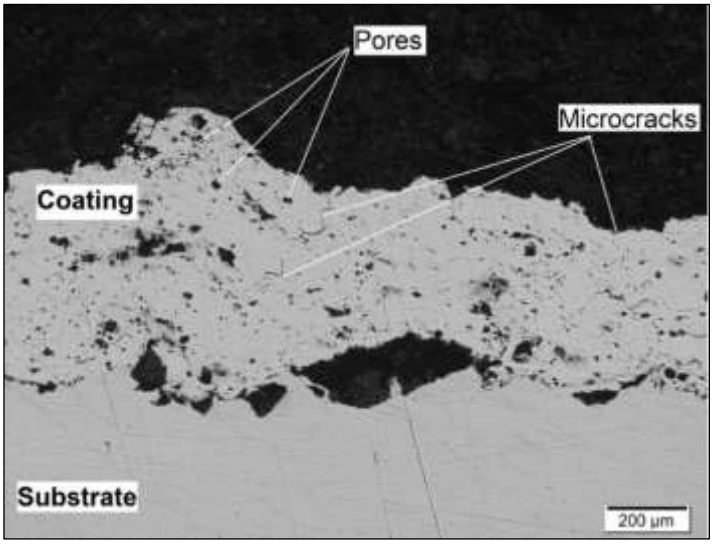


Figure 3: Cross-sectional optical micrograph of a thermal spray coating.
Source: Authors, (2026).

The temperature of incoming droplets can vary from fully molten liquid to a completely solid state, depending on the particle's temperature. Liquid droplets flow freely and cover most of the spaces. The impact velocity of solid or partially melted particles must be sufficiently high to deform plastically upon impact. Consequently, particles with higher velocities undergo greater deformation, resulting in finer void closure. If the impact velocity is modest, solid particles may become lodged in the coating due to subsequent particles approaching later. These unformed particles have poor adhesion and should be further apart from the underlying splat, forming voids [4, 13]. Porosity reflects the fundamental properties of the coating, and typically, lower porosity indicates better coating formation and quality properties. Excessive porosity can reduce strength, abrasion resistance, and corrosion resistance [13]. The pore structure plays a crucial role in determining the performance of coatings in service. Higher porosity levels can compromise certain aspects of coating performance, such as the bond strength between the coating and the substrate.

Porosity diminishes the coating's strength by promoting delamination, cracking, or spallation [13],[14], and properties like density and toughness are significantly impacted by coating porosity [15]. However, in specific scenarios, such as when aiming for coatings with reduced thermal conductivity for wear resistance, porosity might be intentionally introduced to serve as insulation within the coating. Porosity can be controlled in either case, depending on the desired coating characteristics and the appropriate process conditions [4]. According to [13] reported that the HVOF thermal spray process yields coatings with the lowest porosity among commonly used thermal spray processes. This outcome is directly attributed to the high particle velocity and low-temperature flame utilized in the process. The protective effectiveness of coatings against corrosion in various engineering materials heavily relies on coating porosity.

This is particularly critical in systems where a highly reactive (anodic) substrate requires protection from corrosion by means of a cathodic coating [15]. Porous coatings can easily be infiltrated by corrosive agents, allowing them to reach the substrate. In such cases, the coating is prone to early delamination, compromising its protective function. There are two main types of porosity: interconnected and closed (Figure 3). Each type plays a distinct role in determining the corrosion behavior of coatings. Interconnected porosity is significantly more vulnerable to corrosion compared to closed porosity because it offers direct pathways between the substrate and the corrosive environment [4]. Any porosity extending through the thickness exposes a very small anodic area to corrosion, resulting in rapid substrate corrosion due to the high corrosion current density [15]. In [16] investigated Al-5Mg alloy coatings produced via twin-wire arc thermal spray on SS400 steel substrates, employing varied spraying parameters. Porosity analysis followed the ASTM E2109-01 standard, while corrosion performance in 3.5 wt. % NaCl solution was assessed through electrochemical polarization scanning and electrochemical impedance spectroscopy.

Findings revealed that coatings deposited with a spray mode comprising a 300 A current, 5.5 bar pneumatic pressure, and 160 mm standoff distance displayed higher Al and Mg elemental content compared to counterparts fabricated under different parameters. Porosity levels ranged from 7.38 to 11.82%, varying with spray mode. Over time, corrosion resistance varied due to the influence of corrosive products agglomerated on the coating surface. After 240 hours of immersion in 3.5 wt. % NaCl solution, microcracks emerged on the surface of coatings fabricated at 150A spray current, 4.0 bar spray pressure, and 100 mm spray distance, while no cracks were observed on others. XRD analysis indicated that the primary corrosion product on coating surfaces is predominantly Al (OH)3. The study inferred significant enhancement in corrosion mitigation attributed to elevated Al and Mg contents, surface homogeneity, and reduced porosity. Quantifying, characterizing, and minimizing coating porosity are essential for determining and improving the protective efficacy of coatings. Currently, numerous methods are under development for observing and measuring coating pores [14],[15],[17], including image analysis, mercury porosimetry, Brunauer-Emmett-Teller (BET) adsorption measurements, densitometry, high-resolution scanning electron microscopy, helium pycnometry, atomic force probes, contact measurement, ray-three-dimensional scanning imaging, and others.

Among these, the image analysis method, utilizing scanning electron microscope technology, stands out for its simplicity of operation and its ability to accurately quantify the number, shape, size, distribution, and other characteristics of pores. It has proven to be a reliable technique for pore analysis and has been endorsed by the International Organization for Standardization (ISO 26946) as the standard method for evaluating pores in thermal-sprayed coatings [14]. According to [15] utilized dimple grinding combined with optical microscopy to quantify internal porosity across the thickness of nine different electroless nickel-phosphorus (Ni-P) coatings deposited on Mg substrates. Additionally, scanning electron microscopy was employed to verify the identification of pores distinguished by dimpling. The influence of phosphorus (P) content within the Ni-P layers and the overall coating thickness on porosity was analyzed. Porosity reached its peak in coatings with high P content (10 wt% P) and was lowest in those with medium P content (5.2 wt% P). Furthermore, porosity exhibited a consistent decline with the augmentation of coating thickness, ranging from 28 μm to 57 μm. In a study investigating the effects of porosity and thermal gradients introduced during pulsed thermography, in [18] utilized pulsed thermography and a multilayer analytical model to critically assess the thermal diffusivity and inherent porosity in cold-sprayed Ti-6Al-4V coatings on Al6061-T6 substrates.

These coatings are crucial in utilizing cold spray as an additive manufacturing technique for the restoration of high-cost alloy components. The study drew upon experimental data obtained at spray angles ranging from 50◦ to 90◦, illustrating the relationship among the spray angle, porosity within the coating, and the resulting thermal diffusivity. The results revealed a decrease in porosity and an increase in thermal diffusivity as the spray angle increased from 50◦ to 90◦. Infrared thermography measurements of thermal diffusivity showed that the effective thermal diffusivity varied from 3.876% to 26.47% relative to the original coating thermal diffusivity when the spray angle decreased from 90◦ to 50◦. Correspondingly, numerical simulations performed on a partial representative microstructure of the coating illustrated a change ranging from 1.62% to 33.54% from the base coating thermal diffusivity. The analytical model, based on the effective medium theory using the Maxwell–Garnett approximation, forecasted a thermal alteration between 1.62% and 13.05% from the base coating thermal diffusivity. Microscopic analysis-based structural characterization revealed the variation in porosity content in cold spray coating with respect to the spray angle. It was found that the porosity percentage increased from 1.82% to 14.73% when the spray angle varied from 90◦ to 50◦. It was also observed that the porosity increased nonlinearly with the spray angle.

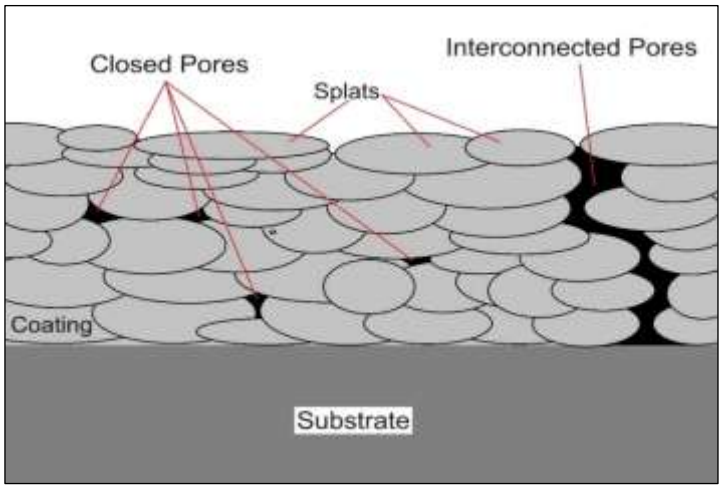


Figure 4: Schematic view of porosity in a thermal spray coating.
Source: Authors, (2026).

In a study investigating the adjustment of deposition parameters such as temperature, bias voltage, and pressure during magnetron sputtering to deposit Cr coatings with varying crystallographic orientations, crystallinity levels, and grain sizes, in [19] observed that during steam oxidation at 800 °C: Coatings deposited at 400 °C, with a bias voltage of –70 V, and at pressures ranging from 0.2 Pa to 0.5 Pa exhibited small grain sizes, high crystallinity, and random orientation. These characteristics contributed to favorable oxidation behavior, primarily due to their dense structure. Coatings deposited with a bias voltage range of 0 V to –40 V showed poor adhesion to the substrate, attributed to the low deposition energy of the plasma. Spallation of the oxide scale or metallic coating occurred during the oxidation test. Coatings deposited at 600 °C, with bias voltages ranging from –100 V to –200 V, and pressures between 2.0 Pa and 3.0 Pa, exhibited a strong (200) texture and large grain sizes. Intrinsic defects such as large grain boundaries and voids expanded into through-thickness cracks and large pores after oxidation. Coatings deposited at temperatures between 24 °C and 200 °C, at a pressure of 1.0 Pa, displayed slight texture and poor crystallinity. Oxygen-rich particles emerged in the un-oxidized coatings beneath the outer Cr2O3 layer after oxidation.

III.2.2 Microcracks

Variances in thermal gradients lead to the formation of cavities and micro-cracks, exacerbated by residual stresses, thermal expansion, and shrinkage during quenching cycles [8]. The roughness of coating surfaces, such as valleys, serves os stress concentration points, fostering crack initiation. Additionally, features like pores and splats in the coating's microstructure contribute to crack branching and deflect crack paths [20] (see Figure 5). The ceramic coating experiences heightened thermal gradient and residual stress due to differences in thermal expansion coefficients between adjacent layers, amplifying thermal mismatch stress within the coating system [21],[22]. In plasma-sprayed thermal barrier coatings (TBC), the bond coat/top coat (BC/TC) interface plays a crucial role due to the formation of the thermally grown oxide (TGO) layer and stress distribution at varying temperatures. Research shows that a smooth interface reduces stress, while rough interfaces increase stress during high-temperature exposure. Additionally, crack propagation mechanisms alter with interface roughness, with higher roughness promoting increased TBC fatigue life by spreading fatigue cracks along the interface and top coat, rather than concentrating them near the bond coat surface [23]. According to [23] devised a discrete element-based model to explore thermal stresses occurring at the interface between layers, attributed to the disparity in coefficients of thermal expansion (CTE).

Their findings suggest that the combination of CTE mismatch and thermally grown oxide (TGO) layer growth induces high tensile stress concentration within the bond coat (BC) layer upon cooling, leading to interface delamination. Conversely, stress reversal occurs in the thermal barrier coating (TC) layer, with increasing interface amplitude. The presence of porosity near crack propagation alters the crack path based on porosity orientation and shape. In regions without porosity, cracks converge to form larger branches, predominantly growing vertically within the ceramic TC. The TGO layer experiences compressive stresses, diminishing as the TC peak region is approached. A thicker TGO layer tends to generate a larger region of compressive stresses, potentially impeding crack propagation into and through the TC layer during the early cooling stages. The extended durability of thermal barrier coatings (TBC) faces a challenge from the infiltration and corrosion caused by calcium-magnesium-aluminum-silicate (CMAS). According to [24] investigated the microstructure and microcrack evolutions of CMAS-corroded TBCs through experiments and numerical modeling. In the study comparative experiments were conducted on TBCs that underwent treatment with and without CMAS at elevated temperatures for various durations.

Comparative experiments revealed that significant volume expansion, stemming from grain dissolution and TC corrosion due to CMAS, is the primary cause of TC spalling. Microcracks enriched with CMAS were observed at the TC/BC interface, resulting from a reaction between CMAS and TGO that degrades the interface. These microcracks act as pathways for CMAS infiltration, hastening the deterioration of CMAS-corroded TBCs. Additionally, the mechanism of microcrack evolution in CMAS-corroded TBCs is linked to the extent of CMAS coverage. When CMAS coverage is minimal, microcracks tend to propagate at the TC/BC interface and TC, leading to localized spalling. Conversely, greater CMAS coverage causes microcracks to extend solely at the TC/BC interface, resulting in the entire TC spalling. A coupled model has been developed, integrating CMAS penetration, TC corrosion and swelling, interfacial degradation, and microcrack propagation, providing insights into the degradation and failure process of CMAS-corroded TBCs as observed in the experiments. The laser surface melting technique, integrated into the deposition process, is renowned for enhancing the corrosion and erosion resistance of coatings. This method achieves heightened erosion resistance by compacting the surface layer. Laser glazing, while reducing surface roughness and enhancing microhardness, also leads to a slight increase in the coating's thermal diffusivity due to reduced surface porosity.

Moreover, laser glazing is effective in halting the spread of molten CMAS through the coating, thereby bolstering its resistance to hot corrosion. This enhancement is attributed to the formation of microcracks, which augment the strain tolerance of the topcoat [22],[25]. According to [25] conducted an analysis on the effects of laser modification post-coating deposition. This modification involves creating a pattern of through-thickness grooves within the coating, followed by surface layer remelting. Their findings concluded that the modified coating exhibited enhanced durability compared to the as-sprayed coating, reducing surface roughness from Sa 9.81 μm to Sa 0.75 μm. This reduction potentially leads to improved tribological flow conditions. Additionally, the modified coating showed an average 8.6 times increase in erosion resistance compared to as-sprayed coatings. This improvement can be attributed to surface densification and reduced influence of poor splat-to-splat adhesion in thermally sprayed coatings. Furthermore, there was an increase in CMAS resistance due to decreased propagation of molten silicates within the coating, along with an 11% enhancement in thermal fatigue resistance.

III.2.3 Unmelted Particles

In the thermal spraying process, uneven heating or irregular feedstock delivery can lead to inadequate melting of the spray particles. The utilization of larger-sized particles has been observed to decrease the temperature of the sprayed particles, resulting in a higher proportion of unmelted particles within the coating layer (shown in Figure 5).

This correlation associates the typical formation of pores in thermal spray processes with the presence of unmelted and semi-melted particles. These unmelted and partially melted droplets often demonstrate limited fluidity and formability, posing challenges in their ability to flatten and spread during deposition. Consequently, voids persist between the splats, which remain unfilled by the unmelted or semi-melted droplets even after the solidification of melted droplets [26, 27]. Unmelted particles represent a minority of the total incident particles (less than 10%), thus the occurrence of large pore formation is infrequent but significant [27]. As per Alebrahin et al. [28], unmelted particles may stem from the deposition of agglomerated submicron or nanosized particles, or from the impact of suspension droplets where not all the water evaporated during flight. The residual water evaporates on the hot substrate, leaving behind unmelted feedstock particles deposited on the substrate surface. In fact, the majority of individual submicron particles probably do not reach the substrate because of their low mass.

The varying sizes of unmelted particles are correlated with the atomization conditions within these spray parameters [29]. Significantly, coatings acquired through twin-wire arc spraving often showcase increased pore content owing to challenges in maintaining a consistent feed rate for feedstock wires compared to feedstock powders. Conversely, the HVOF spraying technique induces considerable plastic deformation of the splats during deposition due to the high particle velocity. Consequently, coatings produced through HVOF spraying typically exhibit dense, robust, and well-bonded characteristics [26]. In their investigation of unmelted particles in plasma spray, Ctibor et al. [30] revealed that typical unmelted particles are predominantly spheroidal and possess a microstructure closely linked to the powder fabrication process prior to spraying. These unmelted globular particles within the lamellar structure of the deposit tend to exhibit isotropic behavior under the influence of stress or other physical factors. Consequently, within the anisotropic structure of the coating, they act as centers of secondary defects, ultimately compromising the functional properties of the coatings.

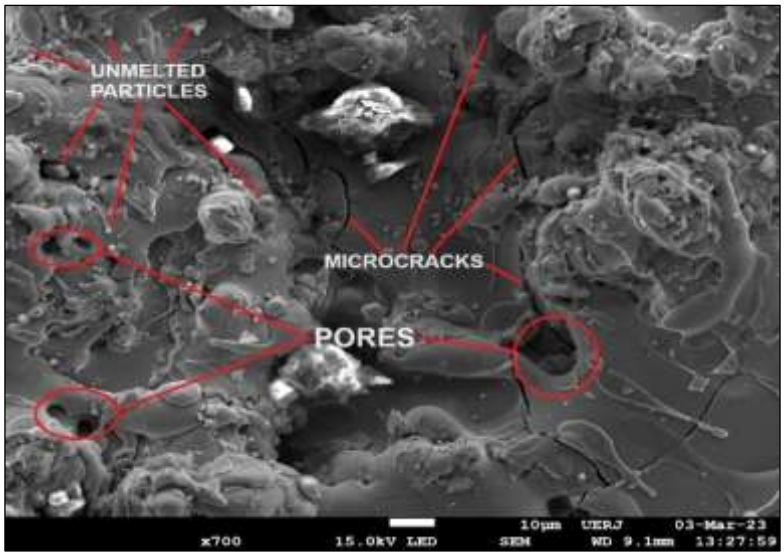


Figure 5: SEM micrograph presenting surface morphology and defects of a thermal sprayed coating.
Source: Authors, (2026).

III.2.4 Oxides

In the high-temperature spraying process, when particles meet the oxidizing atmosphere, they undergo oxidation. Reactive metals like aluminum, nickel, and chromium develop oxide layers during spraying, impacting both microstructure and coating bonding. These oxide layers limit the spread of molten particles upon impact, raising coating porosity and creating friction between the substrate and the splat [4],[26]. In their study, Brossard et al. [31] noted metal particle oxidation during spraying, manifesting in two forms: globular (enclosed within droplets) and as spinel on the outer surface. Subsequently, this oxidation process manifests as a ring-like granular structure within the spray coating. The researchers also attributed the oxidation of certain metals to high temperatures and low oxygen levels. Indeed, the occurrence of oxides in the coating is anticipated due to two factors. Firstly, molten particles are propelled under air pressure during projection. Secondly, each liquid alloy particle is enveloped by an air film during flight, leading to continued oxidation as oxygen becomes trapped upon impact with the target [32]. For metal coatings, a higher degree of oxidation typically corresponds to heightened porosity within the coating layer [25],[32]. Generally, oxidation levels can be reduced by utilizing inert gas, boosting the velocity of spray particles, or minimizing the distance between the spray gun and the substrate [31],[32]. During flame spraying, oxidation of spray particles is unavoidable since combustion requires a blend of oxygen and fuel gases.

Additionally, the oxide content of coatings produced by twin-wire arc spray might increase when compressed air is employed for droplet atomization and propulsion [30]. In HVOF spraying, the supersonic jet speeds up the spray particles, curbing oxide formation and yielding a coating with minimal oxide content. Further reduction in oxide content within HVOF-sprayed coatings can be achieved by lowering the temperature and boosting the velocity of spray particles [26]. According to [33] conducted a study on AlCoCrFeMo high entropy coatings (HECs) produced through three thermal spray techniques: low-pressure cold spray (LPCS), flame spray (FS), and high-velocity oxy-fuel spray (HVOF), aiming to create wear-resistant surfaces. The coatings were examined for their microstructures, phases, mechanical properties, and sliding wear behavior from room temperature up to 350°C. The LPCS, FS, and HVOF coatings initially exhibited a BCC solid solution phase. LPCS coatings lacked oxide inclusions, while FS and HVOF coatings displayed such inclusions due to in-flight oxidation (IFO). HVOF-sprayed coatings exhibited superior hardness due to their finer microstructure, reduced porosity, and fewer oxide inclusions compared to FS and LPCS coatings. The presence of oxide inclusions in FS and HVOF coatings contributed to increased hardness, while LPCS coatings had lower hardness and higher wear rates due to the absence of such inclusions. In terms of

tribological performance, LPCS coatings had lower frictional coefficients compared to FS and HVOF coatings at both tested temperatures. HVOF coatings showed improved wear resistance due to their higher hardness and the formation of a Co-based oxide film. Additionally, the presence of harder and stable Al- and Cr-based oxides provided protection against repetitive stresses during sliding.

III.3 QUANTIFICATION TECHNIQUES

Microstructural characterization plays a pivotal role in understanding and optimizing the performance of thermal spray coatings. These coatings often exhibit complex microstructures with a range of inherent imperfections, including voids, cracks, and unmelted or partially melted particles. Quantitative evaluation of geometrical features—most notably porosity—and detailed analysis of material aspects such as splat morphology, inter-splat interfaces, and phase composition are essential for correlating microstructural attributes with coating properties and in-service behavior. Among these features, porosity is particularly prevalent and has a critical impact on mechanical integrity, thermal conductivity, corrosion resistance, and other functional characteristics. As such, a variety of experimental techniques, including image analysis of cross-sectional micrographs, mercury intrusion porosimetry, and X-ray computed tomography, are commonly employed to measure and characterize porosity with high accuracy. The ability to reliably characterize these features is fundamental to the rational design and control of coating processes, enabling the production of coatings with predictable performance in demanding environments [34]. Nondestructive evaluation (NDE) techniques are widely employed for the detection and characterization of defects in materials and coatings, offering the advantage of preserving the structural integrity of the tested components. Common NDE methods include ultrasonic testing (UT), eddy current testing (ECT), X-ray radiography, and infrared thermography (IRT), among others.

Among these, ultrasonic testing (UT) has seen particularly rapid advancements in recent years, driven by improvements in sensor technology, signal processing, and imaging capabilities. These developments have significantly enhanced the resolution, sensitivity, and reliability of defect detection, making ultrasonic techniques increasingly suitable for the inspection of complex and multilayered coating systems [34]. When ultrasonic waves are transmitted into a sample via a coupling medium, internal defects disrupt their propagation by scattering or reflecting the acoustic energy, whereas regions without defects allow the ultrasound to propagate with minimal attenuation or distortion [34-37]. This contrast in wave behavior forms the basis for defect detection using ultrasonic testing. However, the effectiveness of this technique is limited in coatings with high porosity, as the numerous air-filled voids can scatter or absorb ultrasonic waves, thereby reducing signal clarity and measurement reliability [35-37]. Furthermore, the requirement for physical contact and the use of a coupling agent during measurement pose additional challenges when testing delicate or rough-coated surfaces [34]. Eddy current testing (ECT) operates on the principle of detecting variations in the amplitude and phase of eddy currents induced in a conductive material when subjected to alternating magnetic fields [34]. These eddy currents are influenced by the presence of defects, discontinuities, or changes in material properties, making ECT a valuable technique for surface and near-surface inspection.

However, the accuracy and sensitivity of ECT can be significantly affected by factors such as the electrical conductivity of the material, as well as the geometry, size, and shape of the test specimen. These dependencies necessitate careful calibration and consideration of material-specific parameters to ensure reliable defect detection [38], [39]. Infrared thermography (IRT) has emerged as a widely adopted technique for defect evaluation, owing to its non-contact nature, high efficiency, rapid data acquisition, and capacity for direct and quantitative characterization of subsurface anomalies [40],[41]. These advantages make IRT particularly attractive for real-time and large-area inspection applications. However, current limitations in spatial resolution and the relatively shallow penetration depth restrict its effectiveness for detecting deeply embedded defects. As a result, further research and technological development are required to enhance the resolution and subsurface imaging capabilities of infrared thermography, particularly for applications involving complex or multilayered materials [34]. A variety of experimental techniques are employed for the quantitative measurement and characterization of porosity in coatings, each with distinct advantages and limitations.

Water adsorption (WA) is a simple and cost-effective method; however, it exhibits limited sensitivity in distinguishing between samples with closely similar porosity levels. Mercury intrusion porosimetry (MIP) offers high precision in measuring open porosity but generally underestimates total porosity due to its inability to access closed pores. Helium pycnometry (HP) allows for the measurement of both open and closed porosity—provided the theoretical density of the material is known—but does not yield information on pore size or distribution. Image analysis (IA), commonly conducted on cross-sectional micrographs, enables the evaluation of both open and closed porosity, as well as pore size distribution and morphology. Nevertheless, the accuracy of IA is strongly influenced by spatial resolution and image contrast, and it lacks the ability to distinguish between open and closed pores. Electrochemical testing, although not capable of quantifying total porosity, provides valuable insight into pore connectivity within coatings. This method relies on the infiltration of an electrolyte through interconnected pores, leading to corrosion of the metallic substrate; higher connectivity results in increased corrosion current, offering an indirect measure of porosity interconnectivity [34], [42]. Small-angle neutron scattering (SANS) represents a more advanced technique, capable of resolving component porosities and providing surface area distributions associated with intrasplat cracks, interlamellar pores, and globular pores, including their approximate orientation.

These voids are commonly classified based on aspect ratios—1:1 for globular pores, 1:5 for interlamellar pores, and 1:10 for intrasplat cracks—enabling a more detailed and structural understanding of the coating's porosity network [43],[44]. The use of image processing (IP) techniques presents multiple advantages compared to conventional visual inspection, especially in the context of defect detection and quality control. IP offers enhanced precision and finer detail resolution, making it capable of identifying minor imperfections that may be overlooked by human inspectors. These systems can operate autonomously, significantly reducing the need for manual labor while enabling decisions to be made based on predefined rules and advanced algorithmic analysis. Additionally, IP ensures consistent and repeatable outcomes, which is valuable for tracking changes in defects over time. Remote inspection is another key benefit, as IP allows for the evaluation of areas that are difficult or dangerous to access, thereby increasing safety and eliminating the need for direct human presence. For high-volume production environments, IP is also more economical, lowering costs by increasing inspection throughput. Furthermore, IP systems rely on quantitative analysis, using detailed computations to assess defects and make objective judgments through sophisticated mathematical models. Collectively, these features highlight the clear benefits of IP over visual inspection in terms of reliability, efficiency, and safety [45-50].

According to [51] applied systematic image processing using ImageJ to analyze the microstructure of annealed environmental barrier coatings (EBCs), focusing on crack density, porosity, and phase distribution. Their results showed that REMS-rich coatings (M100 and M75D25) exhibited significant surface cracking (up to 0.58%), attributed to thermal expansion mismatch. In contrast, REDS-rich coatings (M25D75, M10D90, and D100) demonstrated improved microstructural stability, with lower crack coverage (0.04–0.14%) and porosity below 3%, highlighting the beneficial impact of REDS-rich compositions on coating integrity. In the study by [52], the effects of varying heat input on porosity, microstructure, microhardness, residual stress, and thermal behavior of martensitic stainless steel cladding on FGS600-3A ductile cast iron were investigated. Quantitative porosity analysis was performed using digital image processing with the NIS Elements-D software. The results demonstrated that heat input plays a critical role in pore and crack formation. As heat input increased, the porosity rose from 3.41% to 5.56%, with larger, more circular pores propagating between the first and second cladding layers. At lower heat inputs, pore formation was primarily attributed to lack of fusion, balling effect, and thermal contraction gradients.

III.4 ADDRESSING COATING DEFECTS THROUGH PREDICTIVE MACHINE LEARNING MODELS

Machine learning (ML) is a branch of artificial intelligence that enables computers to learn from data and improve their performance on tasks without being explicitly programmed. It involves the use of various algorithms designed to analyze data, identify patterns, and make decisions or predictions. The choice of algorithm depends on factors such as the nature of the problem, the type and number of variables involved, and the most suitable model for the task. ML enhances data handling efficiency and is widely applied across domains to solve complex data-driven problems [53]. Machine learning, is known for its low computational cost and rapid development cycle, is increasingly being applied in material science due to its powerful data processing capabilities and high predictive accuracy. It plays a key role in material detection, analysis, and design, and has been successfully used to discover new materials, predict molecular and material properties, study quantum chemistry, and aid in drug design.

By significantly reducing the need for costly density functional theory (DFT) calculations and repetitive laboratory experiments, machine learning offers an efficient alternative for accelerating scientific discovery [53],[54]. Conventionally, the Taguchi design method has been widely used to identify optimal process parameters and differentiate between major and minor factors influencing coating properties. However, its effectiveness diminishes when complex interactions among numerous parameters affect coating quality [55]. Machine learning (ML) techniques address these limitations by capturing nonlinear relationships and intricate dependencies. Among these, artificial neural networks (ANNs), with their flexible architectures and strong learning capabilities, are particularly effective for predicting and optimizing thermal spraying processes [56]. Artificial Neural Networks (ANNs) are computational models inspired by the structure and function of the human brain. Composed of interconnected artificial neurons organized in layers—typically including an input layer, one or more hidden layers, and an output layer—ANNs are capable of recognizing complex, nonlinear patterns in data.

Each neuron processes input signals by computing a weighted sum, applying an activation function, and producing an output. The weights, analogous to synapses in biological systems, are adjusted through training, allowing ANNs to learn from data and make predictions or classifications. This adaptive learning capability makes ANNs particularly effective in solving intricate problems, such as optimizing conditions in manufacturing processes [55], [57]. In the context of thermal spray processes, artificial neural networks (ANNs) offer a powerful solution for modeling the intricate chemical and thermodynamic reactions that drive coating formation. Given the challenge of acquiring large datasets in this field, shallow ANNs—comprising fewer than three hidden layers—have become a practical choice for effective prediction and analysis in data-scarce scenarios [55]. In a study by [55], an ANN-based regression model was developed to predict key coating properties—such as thickness, porosity, and erosion slurry resistance—based on process parameters like oxygen (O), LPG (L), and air flow (A) rates. Optimized with a single hidden layer of 20 neurons, the model showed strong agreement with experimental results, with validation through SEM imaging, porosity analysis, and mass loss measurements confirming its predictive accuracy. The Backpropagation (BP) neural network is a widely used training algorithm that minimizes prediction error by iteratively adjusting the network's weights and thresholds.

It consists of an input layer, hidden layer(s), and an output layer, operating through forward propagation of signals and backward propagation of errors. In forward propagation, input data undergoes nonlinear transformation through the network to generate an output. If a discrepancy exists between the predicted and actual output, the error is propagated backward to update weights accordingly. To ensure data validity and model effectiveness, input and output features must be preprocessed—commonly through normalization—before training. Through repeated learning, the network converges on weight values that minimize prediction error [58],[59]. The genetic algorithm (GA) is an optimization technique inspired by natural evolution, simulating processes such as gene crossover and mutation to search for optimal solutions. It is particularly effective for solving complex combinatorial problems, often outperforming traditional optimization methods [58]. In the context of BP neural networks, GA can be used to optimize initial weights and thresholds, addressing the variability and instability that arise when these parameters are randomly assigned [58], [60]. The GA-BP neural network combines the strengths of genetic algorithms and backpropagation, offering significant advantages in parameter optimization. By automatically adjusting the network structure and parameters without the need for pre-experiments, this hybrid method enhances the model's generalization ability and improves prediction accuracy [61].

In a study by [61], a 3D transient model combining computational fluid dynamics and surface reaction modeling was developed to analyze flame flow and particle oxidation during the HVOF WC-12Co thermal spraying process. The results revealed that oxide layer thickness varies significantly with particle size, from about 90 Å for 5 µm particles to just 8 Å for 60 µm particles. Using a GA-BP neural network, the researchers optimized key parameters—particle size, oxygen/fuel ratio, and nitrogen flow rate—identifying ideal values for improved coating performance. Experimental validation confirmed that the optimized coating exhibited fewer defects, lower oxide content, and enhanced hardness and wear resistance. Convolutional Neural Networks (CNNs) are a type of feedforward neural network designed to automatically extract meaningful features from raw data using convolutional structures. Inspired by the human visual system, CNNs eliminate the need for manual feature extraction by learning hierarchical patterns directly from the data. They are particularly effective in tasks involving local interactions, such as image recognition and computational materials science.

Key architectural advantages include local connections, which limit each neuron's input to a small region, reducing computational complexity; weight sharing, which minimizes the number of trainable parameters and improves generalization; and pooling layers, which downsample feature maps to retain essential information while discarding redundancy. These features make CNNs efficient, scalable, and well-suited for a wide range of pattern recognition problems [62],[63]. According to [64] developed a deep ensemble convolutional neural network trained on 500 scanning electron microscopy (SEM) images, achieving a test accuracy of 96.0%. Their model demonstrated strong performance across all evaluation metrics, with training, validation, and test accuracies reaching 94.7%, 100%, and 96.0%, respectively. The classification system was particularly effective at distinguishing between different wear patterns, with only minimal confusion between similar mechanisms, resulting in just 4 misclassifications out of 100. This study highlighted how integrating deep learning with materials science can provide a scalable and efficient methodology for tribological analysis and material design.

Physics-Informed Neural Networks (PINNs) are a hybrid machine learning approach that integrates physical laws, typically expressed as partial differential equations (PDEs), directly into the training process of neural networks. By embedding these equations into the loss function through automatic differentiation, PINNs guide the model to produce solutions that are both consistent with the data and compliant with the governing physics of the system. This approach addresses the limitations of traditional machine learning models, such as ANNs and CNNs, which often struggle to generalize to unfamiliar structures. In thermal fluid studies, for example, PINNs have been successfully applied to solve momentum and energy transport equations, accurately predicting particle velocity and temperature. Designed specifically to incorporate any underlying physical law described by nonlinear PDEs, PINNs are especially effective in scenarios with limited data, leveraging known physical principles to enhance model accuracy and reliability [63], [65].

According to [63] developed a hierarchical neural network combining Physics-Informed Neural Networks (PINNs) and Convolutional Neural Networks (CNNs) to predict in-flight particle properties and the performance of NiCr–Cr_3C_2 coatings, including porosity, microhardness, and wear rate. Using a random forest model, they identified spraying distance and particle velocity as the most influential parameters. Their approach achieved prediction accuracies of up to 99%, providing a strong foundation for the development of intelligent HVOF spraying systems. In thermal spray processes, machine learning (ML) offers a powerful way to model complex chemical and thermodynamic reactions. However, limited data availability and the unexplored influence of input variables in ML models present significant challenges.

V. CONCLUSIONS

Thermal spray coatings find widespread application in the fabrication of gas turbine engines, internal combustion engines, bearings, pumping systems, compressor and pipeline equipment, as well as medical implant materials. This technique encompasses a range of cost-effective and dependable methods aimed at safeguarding, repairing, and rejuvenating components, thereby prolonging their operational lifespan even in harsh environments. Defects are inherent in these coatings, emerging from compositional imbalances resulting from metallurgical changes within the structures or irregular surface morphology arising from the layer-by-layer deposition and fusion process on material surfaces. These surface irregularities serve as sites for stress concentration, facilitating crack formation. The presence and nature of defects are greatly influenced by both the feedstock materials and the parameters of the deposition process. Non-dense crystallographic orientation, poor accumulation of large grains, and irregular atomic arrangement in amorphous regions are all detrimental to the compactness of the coatings and negatively affect their oxidation behavior. Although defects are characteristic of thermal spray coatings, machine learning presents a valuable opportunity to advance the control and optimization of thermal spray coatings by modeling the complex chemical and thermodynamic interactions that influence coating quality. Despite the challenges posed by limited data and the underexplored impact of individual process variables, ML enables more accurate prediction of defect formation and supports the development of optimized spraying parameters. As a result, it offers a powerful tool for improving coating performance, reducing defects, and extending the service life of components in demanding environments.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Uirá Montani, Marília Garcia Diniz, Paolo Bettini, Hector Reynaldo Meneses Costa.
Methodology: Uirá Montani, Hector Reynaldo Meneses Costa.
Investigation: Uirá Montani.
Discussion of results: Uirá Montani, Hector Reynaldo Meneses Costa.
Writing – Original Draft: Uirá Montani.
Writing – Review and Editing: Uirá Montani, Marília Garcia Diniz, Hector Reynaldo Meneses Costa.
Resources: Uirá Montani, Marília Garcia Diniz, Paolo Bettini, Hector Reynaldo Meneses Costa.
Supervision: Marília Garcia Diniz, Hector Reynaldo Meneses Costa.
Approval of the final text: Uirá Montani, Marília Garcia Diniz, Paolo Bettini, Hector Reynaldo Meneses Costa.

VII. DECLARATIONS

Data Availability: The data that support the findings of this study are available from the corresponding author upon reasonable request.
Competing Interests: The authors declare that they have no conflict of interest.
Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work, the authors used Chat-GPT in order to improve the readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

VIII. REFERENCES

[1] Sathish, M.; Radhika, N.; Saleh, Bassiouny. Duplex and composite coatings: a thematic review on thermal spray techniques and applications. Metals and Materials International, v. 29, n. 5, p. 1229-1297, 2023.

[2] Singh, Surinder et al. Applications and developments of thermal spray coatings for the iron and steel industry. Materials, v. 16, n. 2, p. 516, 2023.

[3] Zhang, Panpan et al. Microstructures and Wear Resistance of Mo Coating Fabricated by In Situ Laser-Assisted Plasma Spraying. Journal of Thermal Spray Technology, p. 1-13, 2023.

[4] Qadir, Danial et al. A review on coatings through thermal spraying. Chemical Papers, v. 78, n. 1, p. 71-91, 2024.

[5] RigamontI, Daniela et al. Thermal spray to embed optical fibers for the monitoring and protection of metallic structures. Journal of Materials Science, v. 59, n. 27, p. 12812-12829, 2024.

[6] Du, Xinwei et al. Wire arc additive manufacturing from the perspective of remanufacturing: A review of data processing. Journal of Manufacturing Processes, v. 107, p. 385-410, 2023.

[7] Palumbo, Joshua; Chandra, Sanjeev. Additive manufacturing of complex structures and flow channels using wire-arc thermal spray. Journal of Manufacturing Processes, v. 107, p. 459-471, 2023.

[8] Mohanty, Shalini; Gokuldoss Prashanth, Konda. Metallic coatings through additive manufacturing: A review. Materials, v. 16, n. 6, p. 2325, 2023.

[9] Senol, Seren et al. Improved surface quality and fatigue life of high-strength, hybrid particle reinforced (Ti+ B4C)/Al-Cu-Mg metal matrix composite processed by dual-laser powder bed fusion. Procedia Structural Integrity, v. 53, p. 12-28, 2024.

[10] Maleki, Erfan; Bagherifard, Sara; Guagliano, Mario. Correlation of residual stress, hardness and surface roughness with crack initiation and fatigue strength of surface treated additive manufactured AlSi10Mg: Experimental and machine learning approaches. Journal of Materials Research and Technology, v. 24, p. 3265-3283, 2023.

[11] Vaz, Rodolpho F. et al. Cavitation resistance of FeMnCrSi coatings processed by different thermal spray processes. Hybrid Advances, v. 5, p. 100125, 2024.

[12] Arshid, Ehsan; Amir, Saeed; Loghman, Abbas. On the vibrations of FG GNPs-RPN annular plates with piezoelectric/metallic coatings on Kerr elastic substrate considering size dependency and surface stress effects. Acta Mechanica, v. 234, n. 9, p. 4035-4076, 2023.

[13] Nguyen, Tuan-Linh et al. Analysis of the effect of spray mode on coating porosity and hardness when spraying press screws by the high velocity oxy fuel method. EUREKA: Physics and Engineering, n. 6, p. 93-103, 2023.

[14] Li, Bo et al. Study on Porosity of Thermal-Sprayed Commercially Pure Aluminum Coating. Materials, v. 16, n. 19, p. 6612, 2023.

[15] Hu, H. et al. Dimple Grinding Coupled with Optical Microscopy for Porosity Analysis of Metallic Coatings. Micron, v. 178, p. 103593, 2024.

[16] Nguyen, Tuan Van et al. Characterization and corrosion resistance of the twin-wire Arc spray Al-5Mg alloy coating applied on a carbon steel substrate. Journal of Thermal Spray Technology, p. 1-17, 2023.

[17] De Formanoir, Charlotte et al. Healing of keyhole porosity by means of defocused laser beam remelting: Operando observation by X-ray imaging and acoustic emission-based detection. Additive Manufacturing, v. 79, p. 103880, 2024.

[18] Unnikrishnakurup, Sreedhar et al. Exploring thermal dynamics and porosity of cold-sprayed Ti-6Al-4V coatings on Al6061-T6 substrates: A pulsed thermography and numerical modeling approach. International Journal of Thermal Sciences, v. 196, p. 108732, 2024.

[19] Meng, Yan et al. Effect of deposition parameters on characteristics and oxidation behavior of magnetron sputtered Cr coatings. Journal of Nuclear Materials, v. 588, p. 154802, 2024.

[20] Amer, Mohamed et al. Unraveling the Cracking Mechanisms of Air Plasma-Sprayed Thermal Barrier Coatings: An In-Situ SEM Investigation. Coatings, v. 13, n. 9, p. 1493, 2023.

[21] Tao, Shiqian et al. Effect of Cracks on Thermal Shock Behavior of Plasma-Sprayed Thick Thermal Barrier Coatings. Journal of Materials Engineering and Performance, v. 32, n. 11, p. 4998-5014, 2023.

[22] Sokolowski, Pawel et al. The behavior of plasma sprayed thermal barrier coating with laser microtextured bond coat under high temperature testing. Surface and Coatings Technology, v. 453, p. 129095, 2023.

[23] Ferguen, N.; Leclerc, W.; Lamini, E.-S. Numerical investigation of thermal stresses induced interface delamination in plasma-sprayed thermal barrier coatings. Surface and Coatings Technology, v. 461, p. 129449, 2023.

[24] Zhou, Qianqian et al. Microstructure and microcrack evolution mechanism of thermal barrier coatings under CMAS infiltration and corrosion: Experimental and numerical modeling. Corrosion Science, v. 229, p. 111890, 2024.

[25] Ushmaev, Dmitrii; Norton, Andy; Kell, James. Thermally sprayed coatings resistant to environmental degradation: Columnar-like coatings through laser ablation and surface melting approach. Surface and Coatings Technology, v. 460, p. 129394, 2023.

[26] Chen, Tai-Cheng et al. The influence of coating microstructure on the corrosion behavior of Inconel 625 coatings fabricated through different thermal spraying processes. Surface and Coatings Technology, p. 130674, 2024.

[27] De Oca Zapiain, David Montes et al. Calibration of thermal spray microstructure simulations using Bayesian optimization. Computational Materials Science, v. 235, p. 112845, 2024.

[28] Alebrahim, Elnaz et al. Tailoring the porosity level of the suspension plasma sprayed coatings using a dual suspension injection system. Surface and Coatings Technology, v. 478, p. 130401, 2024.

[29] Taghi-Ramezani, Saeid; Valefi, Zia. Effects of Spray Parameters on the YSZ-Alumina Splat Formation During Solution Precursor High Velocity Flame Spraying. Metals and Materials International, p. 1-13, 2023.

[30] ctibor, pavel; roussel, olivier; tricoire, Aurelien. Unmelted particles in plasma sprayed coatings. Journal of the European Ceramic Society, v. 23, n. 16, p. 2993-2999, 2003.

[31] Brossard, S. et al. Study of the microstructure of NiCr splats plasma sprayed on to stainless steel substrates by TEM. Surface and Coatings Technology, v. 204, n. 9-10, p. 1608-1615, 2010.

[32] Boudjit, Sarra et al. Mechanical, Tribological, and Electrochemical Evaluation of NiCrAlMoFe Thermally Sprayed Coatings Subject to Post-treatments. Journal of Materials Engineering and Performance, p. 1-14, 2024.

[33] Patel, Payank et al. Enhanced wear resistance of AlCoCrFeMo high entropy coatings (HECs) through various thermal spray techniques. Surface and Coatings Technology, v. 477, p. 130311, 2024.

[34] Deshpande, Swarnima et al. Application of image analysis for characterization of porosity in thermal spray coatings and correlation with small angle neutron scattering. Surface and Coatings Technology, v. 187, n. 1, p. 6-16, 2004.

[35] Lemlikchi, Safia et al. Ultrasonic characterization of thermally sprayed coatings. Journal of Thermal Spray Technology, v. 28, p. 391-404, 2019.

[36] Meng, Jiajian et al. Mechanical properties measurement of chromium coatings based on laser ultrasonic technology in high-temperature conditions. Mechanical Systems and Signal Processing, v. 228, p. 112421, 2025.

[37] Zawischa, Martin; Makowski, Stefan; Toma, Filofteia-Laura. Fast and Nondestructive Mechanical Characterization of Thermally Sprayed and Laser Cladded Coatings: Automated Surface Acoustic Wave Spectroscopy as a New Tool for Quality Control and Research. Journal of Thermal Spray Technology, v. 34, n. 2, p. 928-938, 2025.
[38] He, Dongfeng; Kusano, Masahiro; Watanabe, Makoto. Detecting the defects of warm-sprayed Ti-6Al-4V coating using Eddy current testing method. Ndt & E International, v. 125, p. 102565, 2022.

[39] Tytko, Grzegorz et al. High frequency eddy current method in inspection of aluminide coatings integrity after simulating service loads. Measurement, v. 252, p. 117356, 2025.

[40] Wei, Jinfeng et al. Quantitative characterization of interfacial defects in thermal barrier coatings by long pulse thermography. Coatings, v. 12, n. 12, p. 1829, 2022.

[41] CernuschI, Federico; Bison, Paolo. Thirty Years of Thermal Barrier Coatings (TBC), Photothermal and thermographic techniques: Best practices and lessons learned. Journal of Thermal Spray Technology, v. 31, n. 4, p. 716-744, 2022.

[42] Andreola, Fernanda; Leonelli, Cristina; Romagnoli, Marcello. Techniques Used to Determine Porosity. American Ceramic Society Bulletin, v. 79, n. 7, p. 49-52, 2000.

[43] Allen, Andrew J. et al. Microstructural characterization of yttria-stabilized zirconia plasma-sprayed deposits using multiple small-angle neutron scattering. Acta Materialia, v. 49, n. 9, p. 1661-1675, 2001.

[44] Kulkarni, A. et al. Comprehensive microstructural characterization and predictive property modeling of plasma-sprayed zirconia coatings. Acta Materialia, v. 51, n. 9, p. 2457-2475, 2003.

[45] Mehrali, Hossein et al. Detection of coating defects in high gas turbine blades by image processing. Signal, Image and Video Processing, v. 19, n. 5, p. 360, 2025.

[46] Deng, Liwei; Guo, Yangang; Chai, Borong. Defect detection on a wind turbine blade based on digital image processing. Processes, v. 9, n. 8, p. 1452, 2021.

[47] Aust, Jonas et al. Automated defect detection and decision-support in gas turbine blade inspection. Aerospace, v. 8, n. 2, p. 30, 2021.

[48] Rizk, Patrick et al. Wind turbine blade defect detection using hyperspectral imaging. Remote Sensing Applications: Society and Environment, v. 22, p. 100522, 2021.

[49] Aminzadeh, Ahmad et al. Non-contact inspection methods for wind turbine blade maintenance: techno–economic review of techniques for integration with industry 4.0. Journal of Nondestructive Evaluation, v. 42, n. 2, p. 54, 2023.

[50] Rao, Yizhuo et al. Wind turbine blade inspection based on unmanned aerial vehicle (UAV) visual systems. In: 2019 IEEE 3rd conference on energy internet and energy system integration (EI2). ieee, 2019. p. 708-713.

[51] Wang, Keyu et al. An in-depth exploration of microstructural characteristics and phase evolution in RE-silicates (RE= Er0. 25Tm0. 25Yb0. 25Lu0. 25) multiphase environmental barrier coatings utilizing digital image processing techniques. Journal of the European Ceramic Society, p. 117276, 2025.

[52] Altay, Meryem; aydin, hakan; karşi, adem. effect of heat input on martensitic stainless-steel laser clad characteristics on ductile cast iron. welding in the world, p. 1-17, 2025.

[53] Mahesh, Batta et al. Machine learning algorithms-a review. International Journal of Science and Research (IJSR). [Internet], v. 9, n. 1, p. 381-386, 2020.

[54] Wei, Jing et al. Machine learning in materials science. InfoMat, v. 1, n. 3, p. 338-358, 2019.

[55] Singh, Vikrant et al. Predictive modeling of HVOF-sprayed TiC coating: an ANN-based approach for coating properties estimation. International Journal on Interactive Design and Manufacturing (IJIDeM), v. 19, n. 3, p. 1709-1720, 2025.

[56] Gui, Longen et al. Prediction of in-flight particle properties and mechanical performances of HVOF-sprayed NiCr–Cr3C2 coatings based on a hierarchical neural network. Materials, v. 16, n. 18, p. 6279, 2023.

[57] Kumar, Suresh S. et al. Numerical analysis of thermal spray coatings using artificial neural networks (ANN) overview. International Journal on Interactive Design and Manufacturing (IJIDeM), p. 1-16, 2024.

[58] Ye, Dongdong et al. Prediction and analysis of the grit blasting process on the corrosion resistance of thermal spray coatings using a hybrid artificial neural network. Coatings, v. 11, n. 11, p. 1274, 2021.

[59] Zhou, Kai et al. Fault diagnosis of spraying workshop based on BP neural network. In: Journal of Physics: Conference Series. IOP Publishing, 2021. p. 032074.

[60] JIang, Qin et al. Application of BP neural network based on genetic algorithm optimization in evaluation of power grid investment risk. IEEE Access, v. 7, p. 154827-154835, 2019.

[61] Li, Siyu et al. Evaluation and Analysis of Particle Oxidation of HVOF Thermal Spraying Based on GA-BP Neural Network Algorithm. Journal of Thermal Spray Technology, p. 1-24, 2025.

[62] Li, Zewen et al. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, v. 33, n. 12, p. 6999-7019, 2021.

[63] Gui, Longen et al. Prediction of in-flight particle properties and mechanical performances of HVOF-sprayed NiCr–Cr3C2 coatings based on a hierarchical neural network. Materials, v. 16, n. 18, p. 6279, 2023.

[64] Paprk, Tae-Jun et al. Deep Learning-Enhanced Wear Analysis of Fe-Cr Composite Coatings for Automotive Brake Applications Under Euro 7 Requirements. Sahn and Oh, Yoon-Suk, Deep Learning-Enhanced Wear Analysis of Fe-Cr Composite Coatings for Automotive Brake Applications Under Euro, v. 7, 2025.

[65] Bobzin, K.; Heinemann, H.; Dokhanchi, A. Physics-Informed Neural Networks for Predicting Particle Properties in Plasma Spraying. In: International Thermal Spray Conference. ASM International, 2024. p. 452-458.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 188-200. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2854

RESEARCH ARTICLE OPEN ACCESS

PREDICTING CUSTOMER SATISFACTION THROUGH LAYERED NLP TECHNIQUES: TRADITIONAL MEETS DEEP LEARNING

Mohammed Bachir Mahdjoub*[1], Fatima Zohra Laallam[2] and Messaoud Mezati[3]

[1]Mohamed Khider University, Algeria
[2, 3]Kasdi Merbah University, Algeria

[1]http://orcid.org/0000-0001-9792-1831, [2]http://orcid.org/0000-0003-2875-3576,[3]http://orcid.org/0009-0001-1996-5625

Email: *bachir.mahjoub@univ-biskra.dz, laallam.fatima_zohra@univ-ouargla.dz, mezati.messaoud@univ-ouargla.dz

ARTICLE INFO

Article History
Received: October 25, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Customer Satisfaction Prediction,
Natural Language Processing,
Deep Learning,
Machine Learning,
Sentiment Classification,
CRM.

ABSTRACT

This study presents a robust hybrid artificial intelligence framework designed to predict customer satisfaction based on online reviews a critical capability for enhancing business performance and optimizing customer relationship management (CRM) practices. We propose a unified architecture with high precision for sentiment classification, utilizing the "All Beauty (2023)" dataset from Amazon Reviews. The approach synergistically combines conventional machine learning (ML) algorithms with state-of-the-art deep learning (DL) techniques. Although standalone ML and DL models have demonstrated promising outcomes in sentiment analysis, they often fall short in addressing intrinsic data challenges such as class imbalance and linguistic subtleties. To overcome these limitations, we introduce HybridStack, a novel ensemble architecture that integrates diverse ML and DL models to significantly improve classification accuracy. The methodology encompasses rigorous text preprocessing, a dual-path feature extraction pipeline (TF-IDF for ML models, and embedding techniques for DL models), and a meta-classifier to consolidate predictions into three categories: dissatisfied, neutral, and satisfied. To address class imbalance, we employ weighted loss functions alongside oversampling strategies. The proposed HybridStack model achieved an outstanding and unparalleled F1-score of 0.6889 and an accuracy of 0.8593, substantially outperforming all baseline individual models. This hybrid approach highlights the benefits of algorithmic synergy and demonstrates its effectiveness in generating more granular and actionable insights, thereby supporting the development of more refined CRM strategies and fostering stronger customer relationships.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Automated analysis of customer satisfaction from online reviews represents a crucial challenge in data science and artificial intelligence, as it directly contributes to enhancing sales strategies and customer relationship management (CRM) effectiveness. Despite significant advances in natural language processing (NLP), machine learning (ML), and deep learning (DL), existing methods still struggle to deliver robust and generalizable performance in real-world, unbalanced datasets. Current studies mainly focus on either traditional ML models or advanced neural networks, often neglecting the complementary strengths that could arise from their integration. In this context, the present work proposes a novel hybrid AI framework that unifies classical ML algorithms (e.g., Logistic Regression, Naive Bayes) with state-of-the-art DL architectures (LSTM, GRU, CNN) and transformer-based models (BERT) within a stacked ensemble learning strategy. This hybridization enables the system named HybridStack to exploit the interpretability and efficiency of traditional classifiers alongside the representational depth of neural and transformer-based models. The proposed pipeline operates in parallel, leveraging TF-IDF features for conventional models and embedding representations for deep architectures, before aggregating their predictions through a meta-learner to enhance overall precision and robustness.

Evaluated on the "All Beauty (2023)" Amazon Reviews dataset, this framework demonstrates improved predictive accuracy and stability compared to single-model baselines. The novelty of this work lies in its systematic integration of heterogeneous learning paradigms within a unified architecture, addressing key challenges such as linguistic complexity, feature diversity, and class imbalance. This contribution advances the state of research toward more adaptive and generalizable satisfaction prediction systems in practical CRM applications.

II. THEORETICAL REFERENCE

II.1 CRM AND CUSTOMER SATISFACTION

Customer relationship management (CRM) is fundamentally a strategic business approach focused on managing interactions with current and potential customers in order to stimulate acquisition, promote loyalty and support long-term retention. Its main objective is to improve customer satisfaction through personalised engagement, a deep understanding of their expectations and the establishment of lasting relationships that promote increased customer lifetime value (Figure 1).

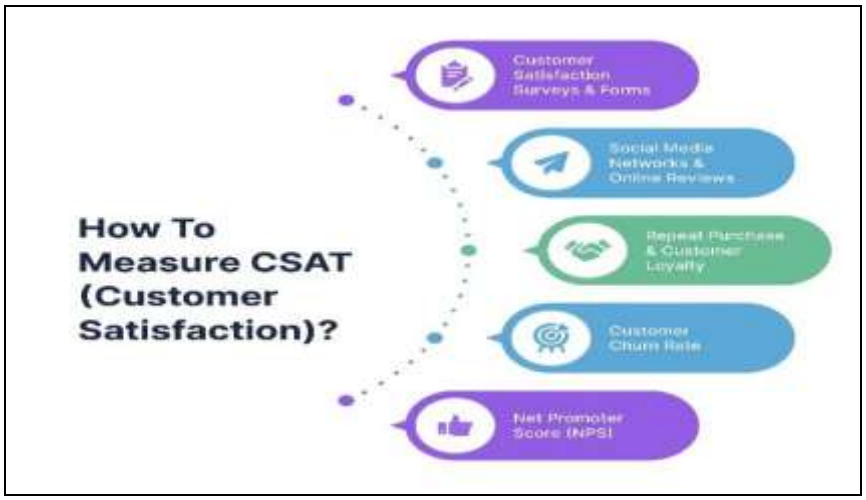


Figure 1: Customer Satisfaction Measurement Techniques.
Source: Authors, (2026).

CRM integrates technology, data and business strategies to systematically collect, process and analyse customer information, enabling organisations to provide more personalised and effective services. Several definitions reflect the multifaceted nature of CRM. [1] defines it as the strategic deployment of information, processes, technologies and human resources throughout the customer lifecycle, including marketing, sales, service and support. [2] presents CRM as a proactive strategy aimed at understanding and anticipating customer needs. Similarly, [3] and [4] emphasise the role of CRM in identifying profitable customer segments, implementing value-driven interactions and developing profitable strategies to drive profitability. [5] characterises CRM as a continuous process of identifying customer value and personalising interactions in order to exceed expectations. [6] further describe CRM as a fundamental business strategy that harmonises internal operations and extends organisational networks to create value through the effective use of high-quality customer data and advanced information systems. In the rapidly expanding e-commerce sector, customer satisfaction is essential for commercial success and market survival [7]. Advanced analytical tools are indispensable for effectively analysing the vast amounts of data generated by customer feedback, which directly influences the success of products and companies [8].

II.2 PREDICTING CUSTOMER SATISFACTION IN THE ERA OF MACHINE LEARNING AND DEEP LEARNING

Sentiment analysis, a key text mining technique, is widely used to identify and extract subjective information from various sources, helping businesses understand customer opinions, assess their preferences, and improve their purchasing decisions [8]. The integration of machine learning and deep learning models, such as BERT, has significantly improved the effectiveness of sentiment analysis and enabled highly accurate predictions of customer satisfaction by analysing reviews and integrating them with e-commerce data [8-11]. This data-driven insight is essential not only for improving product quality and operational efficiency, but also for enhancing the overall customer experience, optimising conversion rates, and fostering sustainable growth in the competitive e-commerce landscape [12]. The analysis of complex textual data, such as customer reviews, relies heavily on the capabilities of artificial intelligence paradigms such as machine learning (ML) and deep learning (DL). Traditional ML models, including logistic regression and the naive Bayes model, have consistently provided robust and interpretable solutions for text classification tasks, often leveraging statistical features such as TF-IDF to quantify the importance of words. At the same time, deep learning models, such as long short-term memory (LSTM), gate recurrent units (GRU), and convolutional neural networks (CNN), have revolutionised natural language processing with their ability to automatically learn complex semantic and contextual representations through integrations, often enhanced by powerful transformer-based architectures such as BERT. These two approaches offer distinct advantages for discerning patterns and sentiments within large datasets, thereby contributing significantly to the automated prediction of customer satisfaction (see Figure 2).

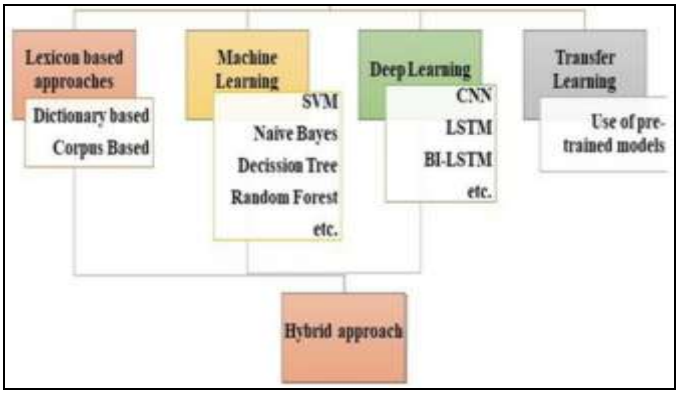


Figure 2: Different Approaches for Text Analysis in Customer Satisfaction Prediction.
Source: Authors, (2026).

II.3 STUDY EXISTING APPROACHES

Traditional Machine Learning (ML) and Deep Learning (DL) models each possess significant capabilities but also distinct limitations when applied to complex, noisy, and imbalanced textual data typical of e-commerce environments. ML models, though interpretable and efficient, often fail to capture the deep semantic and contextual dependencies present in natural language. Conversely, DL models such as LSTM, GRU, and CNN can learn intricate linguistic patterns but tend to act as "black boxes," with reduced interpretability and a higher risk of overfitting especially when faced with underrepresented sentiment classes like dissatisfied customers. These challenges highlight the need for a hybridized analytical framework capable of exploiting the complementary strengths of both paradigms while mitigating their respective weaknesses. The proposed HybridStack approach fulfills this requirement by integrating diverse ML and DL models within an ensemble structure, where a meta-classifier synthesizes their outputs. By combining models specializing in statistical feature extraction (e.g., Logistic Regression, SVM) with those capturing deep semantic meaning (e.g., LSTM, CNN, BERT), HybridStack enhances robustness, accuracy, and generalization.

This fusion enables more reliable sentiment classification and superior detection of minority sentiment classes, directly supporting actionable insights for Customer Relationship Management (CRM) systems. The rapid expansion of e-commerce, driven by Artificial Intelligence (AI) and ML, has intensified research in predictive analytics applications such as demand forecasting, customer segmentation, and fraud detection [13-17]. Sentiment analysis and satisfaction prediction have become central to this effort, leveraging models like SVM, Naïve Bayes, BiLSTM, and BERT to extract subjective opinions from large volumes of customer reviews [8], [9], [18]. Empirical studies across platforms such as Amazon and Jumia confirm the value of these methods [10], [19]. However, standalone models still face persistent obstacles linked to linguistic variability, class imbalance, and noisy data, which hinder their generalizability. To overcome these issues, hybrid and ensemble learning approaches have gained prominence for their capacity to deliver higher predictive accuracy and resilience [12].

Research demonstrates that combining algorithms for instance, SVM with Logistic Regression [13], or multiple ensemble classifiers for churn and delivery prediction [16], [20], [21], [22] achieves superior outcomes. Advanced Natural Language Processing (NLP) techniques have also improved sentiment extraction, topic modeling, and satisfaction forecasting [23-25]. Moreover, the integration of ML with Business Intelligence (BI) and visualization tools [26] supports more informed, data-driven decision-making while ongoing developments in explainable and fairness-aware AI address transparency and bias concerns. Collectively, this growing body of work underscores the transition toward integrative and adaptive analytics frameworks in e-commerce. By uniting interpretability, statistical precision, and deep contextual learning, hybrid systems such as HybridStack represent the most accurate, scalable, and reliable approach for multi-class sentiment prediction. These architectures not only enhance predictive performance but also enable personalized customer experiences, ethical AI implementation, and sustainable business growth in the evolving digital marketplace.

II.4 SCIENTIFIC AND CONTEXTUAL RELEVANCE OF THE RESEARCH

This observation highlights the current relevance of research focused on combining machine learning and deep learning. In a commercial environment increasingly dominated by e-commerce, companies generate massive volumes of textual data every day from reviews, comments and customer interactions. Effectively exploiting this information requires models that are capable of both understanding the linguistic and emotional subtleties of texts and producing reliable, explainable and generalisable predictions. This issue is all the more crucial given that customer satisfaction is a major strategic lever for competitiveness, loyalty and value creation. Thus, the ability of artificial intelligence to accurately predict this satisfaction represents a major scientific, technological and managerial challenge.

II.5 NOVELTY AND POSITIONNING OF RESEARCH

In this perspective, the novelty of this work lies in the design of a hierarchical hybrid approach, combining traditional machine learning (ML) methods and deep learning (DL) models applied to natural language processing. The proposed approach is based on an overall strategy inspired by the HybridStack concept, which aggregates the predictive outputs of several heterogeneous models to feed a meta-classifier. This hybridisation makes it possible to exploit the complementarity between the statistical recognition of ML models (such as SVM, logistic regression, etc.) and the deep semantic understanding of DL models (such as LSTM, GRU, CNN and BERT).

The objective is to simultaneously improve the robustness, accuracy and generalisation capacity of the customer satisfaction prediction system, while maintaining better interpretability of the results. Furthermore, the integration of hierarchical language processing is an original contribution. By structuring textual analysis on several levels lexical, syntactic, semantic, and pragmatic this approach offers a more detailed view of the emotions, intentions, and judgements expressed in customer reviews. It also allows for better detection of minority classes, particularly dissatisfied customers, whose recognition is essential for CRM. Thus, this research builds on recent work on ML and NLP, while offering a distinct and innovative contribution: a hybrid and hierarchical predictive framework that reconciles interpretability, semantic depth and operational performance. This framework aims to fill a gap identified in the literature, namely the lack of integrative models combining symbolic learning and deep representation for the automated prediction of customer satisfaction in e-commerce.

Table 1: Summary of Related Works on E-commerce Customer Behavior and Prediction.

Article	Methodology/Algorithms Used	Key Findings/Contributions
I. Deep Learning Models		
[11]	RNNs, LSTM with SGD	Superior accuracy in dynamic pricing; balanced profit/satisfaction.
[27]	CNNs (VGG16, ResNet50, InceptionV3)	CNNs improve profit prediction (high accuracy).
[10]	Keras DL framework, Sequential model, Adam optimizer	Adam optimizer excels in customer satisfaction prediction.
[9]	ML: BERT model (fine-tuned)	BERT accurately forecasts customer sentiment (96%).
[22]	ML: PIVD using CNN in three steps	PIVD improves image vulnerability detection (high F1).
II. Traditional Machine Learning Models		
[14]	ML: Decision Trees, Random Forest (using TIPI test for consumer profiles)	ML predicts consumer behavior from personality.
[16]	ML: SVM, Random Forest, Decision Tree	SVM predicts churn (83.45%) and delivery.
[28]	ML: Association rules via Frequent-Pattern-Growth algorithm	High probability of next product purchase via recommendations.
[20]	ML: Seven traditional classification algorithms including Logistic Regression, Random Forest, Naïve Bayes	ML predicts online banking satisfaction/frustration.
[26]	BI tools incorporating Data mining, Data Analytic, Data Visualization and ML; ABCD framework	BI/ML aid decision-making and competitiveness.
III. Hybrid/Fusion Machine Learning Approaches		
[13]	ML Fusion Algorithm: Logistic Regression + SVM	Fused ML models offer higher prediction accuracy.
[8]	ML: SVM, Naïve Bayes, DL (bidirectional LSTM, BERT)	Sentiment analysis/forecasting enhances BI/decisions.
[29]	ML: RF, GBDT, XGBOOST; DL: RNNs, LSTM	ML/DL accurately forecast purchase behavior (similar performance).
[18]	Advanced ML techniques: DL, Reinforcement Learning, NLP	ML frameworks boost customer engagement/LTV/growth.
[10]	Multimodal Learning Framework (including unimodal comparison), Adam optimizer	Multimodal learning improves satisfaction/efficiency.
[15]	Big Data Algorithms, Text Mining Technology, Empirical analysis (SEM after correction)	Online interaction impacts satisfaction; personalized recs need ethics.

Source : Authors (2026)

III. MATERIALS AND METHODS

This section presents a hybrid AI framework designed to analyze and enhance customer satisfaction using the Amazon Reviews dataset. The system integrates Machine Learning (ML), Deep Learning (DL), and transformer-based models to deliver a robust sentiment analysis solution. Text preprocessing involves cleaning (removal of symbols, punctuation, and numbers), tokenization, and normalization, followed by numerical conversion through vectorization to prepare data for computation. The architecture adopts a two-stream design. The first stream employs TF-IDF vectorization to extract statistical text features, which are then classified using traditional ML algorithms such as Logistic Regression, Multinomial Naive Bayes, XGBoost, and LinearSVC, chosen for their interpretability and stability as baselines. The second stream applies DL techniques leveraging word embeddings to capture semantic context. Models include LSTM, GRU, and CNN, with pre-trained GloVe embeddings and fine-tuned BERT transformers providing deep contextual understanding through attention mechanisms. Predictions from both streams are integrated through a stacked ensemble meta-classifier, forming the HybridStack. This ensemble combines the complementary strengths of ML and DL models, improving accuracy, robustness, and the classification of sentiment categories (‘dissatisfied,’ ‘neutral,’ ‘satisfied’). Finally, an evaluation module assesses performance using Macro F1-Score, Accuracy, Precision, and Recall, with special focus on minority classes like ‘Dissatisfied.’

Confusion matrices visualize misclassifications, and McNemar's test validates statistical significance. The architecture balances precision, interpretability, and scalability, offering a deployable solution for real-world customer feedback analytics.

III.1 DATA SOURCE AND COLLECTION

The dataset utilized in this study is the Amazon Customer Reviews Dataset, an extensive collection publicly available via the Amazon Web Services (AWS) Open Data Registry. This corpus encompasses a vast array of customer reviews, spanning multiple international marketplaces and diverse product categories. Each review is enriched with valuable metadata, including a unique product identifier (ASIN), reviewer ID, review title and body, a numerical star rating (1 to 5), and the timestamp of submission. For this research, we specifically focused on a curated subset: the Amazon Reviews All Beauty (2023) dataset. This dataset, stored in file "All_Beauty.jsonl", contains customer reviews specifically within the "All Beauty" product domain. This particular domain was selected to ensure a consistent context for product satisfaction, allowing for a focused analysis of beauty-related reviews. The All Beauty category was chosen to facilitate:

- Domain-Specific Textual Expression: Allowing for the capture of unique vocabulary and review writing styles prevalent within the beauty product sector.
- Targeted Customer Expectations: Providing insights into satisfaction drivers specific to beauty products, which can differ significantly from other categories.
- Specialized Model Development: Enabling the training of models highly optimized for predicting satisfaction in this specific retail segment.

The "All_Beauty.jsonl" dataset comprises reviews from the specified category, providing a rich source of information for modeling customer satisfaction. The star rating, from 1 to 5, serves as the primary indicator for satisfaction, allowing for both binary (satisfied vs. not satisfied, typically based on ratings of 4 or 5) and potentially multi-class satisfaction modeling.



Figure 3: Sample Customer Review from the Amazon Reviews All Beauty Dataset.
Source: Authors, (2026).

The Amazon Reviews dataset, and specifically the "All Beauty" subset, stands as one of the largest and most detailed open-access corpora for customer feedback within a specialized domain. It offers a rich combination of textual, categorical, temporal, and numerical features. Its focused coverage on a single product domain, coupled with detailed review content and structured metadata, makes it ideally suited for natural language processing (NLP) tasks such as sentiment classification and user profiling tailored to the beauty industry. Furthermore, the dataset reflects real-world challenges such as class imbalance and noise inherent in user-generated content, making it valuable for simulating realistic AI deployments. Its scale and granularity also enable the training of a wide range of models, from traditional machine learning algorithms to sophisticated neural architectures, including the fine-tuning of transformer-based language models like BERT and GPT for domain-specific applications.

III.2 FEATURE SELECTION

To analyze customer satisfaction in the beauty product domain, several essential features were extracted from the Amazon Reviews – All Beauty (2023) dataset. The text field contains full customer reviews and serves as the primary input for all NLP tasks such as sentiment classification and satisfaction prediction, capturing nuanced emotional expressions and user experiences. The rating feature, ranging from 1 to 5 stars, provides the ground-truth labels for supervised learning, enabling both binary and multi-class sentiment categorization. The title field offers concise summaries of opinions and acts as auxiliary textual input, improving model accuracy and supporting summarization or response generation. The timestamp attribute records the review submission time, allowing for the analysis of temporal satisfaction trends, product lifecycle patterns, and seasonal variations within the beauty sector. The asin (Amazon Standard Identification Number) uniquely identifies each product, enabling product-level sentiment analysis, satisfaction comparison, and recommendation insights. Together, these features form a multi-view dataset combining textual and structured attributes. This integration enhances semantic understanding, contextual awareness, and trend-based prediction, allowing models to learn not only what customers express but also when and about which products crucial for accurate and actionable customer satisfaction modeling in the beauty domain.

III.3 PREPROCESSING

The feature engineering and preprocessing strategy for the **Amazon Reviews – All Beauty (2023)** dataset was carefully adapted to ensure compatibility with both traditional ML and modern DL sentiment analysis models. The dataset structure was standardized by renaming key fields to align with the beauty domain: *review_body → text*, *star_rating → rating*, *review_title → title*, *review_date → timestamp*, and adding *asin* as a unique product identifier. These modifications preserved the dataset's integrity while ensuring semantic alignment for satisfaction modeling.

Text Normalization: Comprehensive normalization procedures were applied to the *text* and *title* fields to ensure uniformity and minimize lexical variability. All text was converted to lowercase, extraneous whitespace was trimmed, and punctuation was removed to reduce noise. Repetitive character sequences (e.g., "soooo") were standardized to maintain consistency and control token distribution.

Noise Filtering: Regular expressions were used to remove HTML tags, escape sequences (e.g., \n, \xa0), and non-alphanumeric symbols that could distort tokenization or interfere with semantic learning. This step ensured clean, well-structured text inputs by eliminating web-scraping artifacts common in Amazon datasets.

Stopword Removal: Stopwords were removed using the NLTK English stopword list to reduce feature dimensionality and enhance discriminative vocabulary for classical ML models (e.g., TF-IDF). However, this step was intentionally skipped for deep learning and transformer architectures to preserve syntactic and contextual relationships critical for attention-based modeling.

Tokenization Strategy: Tokenization was customized to match model requirements. For LSTM-based models, word-level tokenization using Keras Tokenizer produced integer sequences compatible with GloVe embeddings, preserving sequence semantics. For transformer models (e.g., BERT), subword tokenization via Hugging Face AutoTokenizer was implemented, maintaining alignment with the model's pretraining scheme (*bert-base-uncased*). This approach ensures robust handling of rare and out-of-vocabulary words.

Label Encoding: The raw 1–5 rating scale was consolidated into a three-class satisfaction scheme to reduce label sparsity and support balanced multi-class sentiment classification—typically mapping low ratings to dissatisfied, mid-range to neutral, and high ratings to satisfied. Collectively, these preprocessing and encoding strategies create a **standardized, noise-free, and semantically enriched dataset**, enabling consistent input for diverse ML/DL models. The resulting framework supports accurate, interpretable, and scalable **customer satisfaction prediction** within the beauty domain.

Table 2: Label Encoding Scheme for Customer Satisfaction Ratings.

Raw Rating	Encoded Class	Satisfaction Level
1–2	0	Dissatisfied
3	1	Neutral
4–5	2	Satisfied

Source : Authors (2026)

This transformation merges adjacent rating levels with similar semantic meaning, reflecting typical business analytics categorizations of customer satisfaction. It simplifies model training and evaluation while preserving important distinctions among dissatisfaction, neutrality, and satisfaction.

III.4 DATA SPLITTING AND CLASS BALANCE

The dataset was divided into training (70%), validation (15%), and test (15%) subsets using stratified sampling to preserve the original class distribution. This ensured that each subset accurately reflected real-world satisfaction proportions, minimizing sampling bias and supporting reliable model generalization. An exploratory class distribution analysis revealed a strong imbalance, with most reviews concentrated in 4- and 5-star categories, while 1-, 2-, and 3-star reviews were scarce. This skew mirrors real e-commerce behavior, where positive feedback dominates and negative reviews—though less frequent contain vital insights for customer service improvement. An exploratory analysis of the satisfaction labels (star_rating mapped to classes) showed a markedly imbalanced distribution(figure 4):

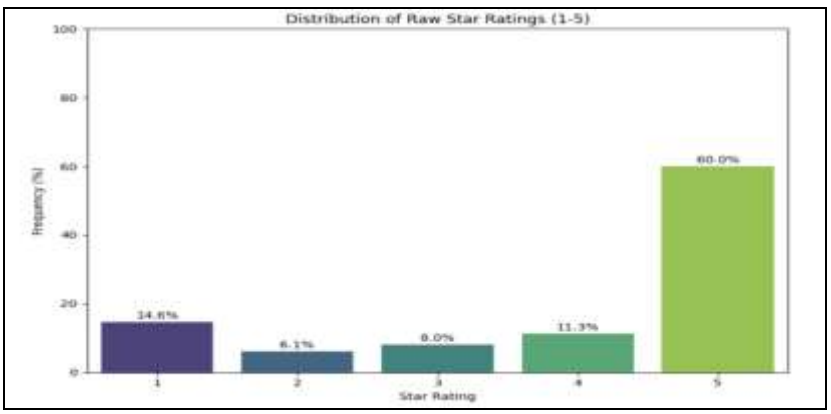


Figure 4: Distribution of Customer Satisfaction Levels.
Source: Authors, (2026).

To mitigate class imbalance, two complementary strategies were implemented. Weighted loss functions assigned higher penalties to minority classes, improving recall for "Dissatisfied" reviews and preventing overfitting to majority classes. This was achieved through the class_weight parameter in TensorFlow/Keras and customized loss weighting in PyTorch.
Additionally, oversampling techniques such as Random Oversampling and SMOTE were selectively applied to the training data, generating synthetic minority samples to balance class frequencies while preventing data leakage. Combined with stratified batch sampling, these methods improved model robustness and minority-class detection These measures enhanced the framework's sensitivity to dissatisfied customers, ensuring that negative feedback signals were effectively captured. By correcting bias toward the "Satisfied" class, the final models achieved a more realistic, balanced, and business-relevant performance, crucial for actionable customer satisfaction prediction in e-commerce (figure 5).

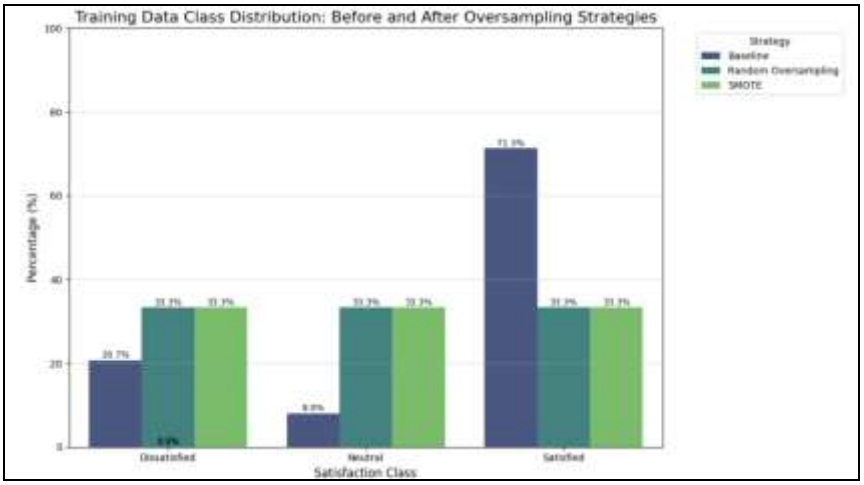


Figure 5: Comparison of Customer Satisfaction Class Distribution with Imbalance Mitigation Strategies.
Source: Authors, (2026).

III.5 PROPOSED HYBRID CLASSIFICATION ARCHITECTURE

This study employs a robust hybrid modeling approach, combining traditional machine learning (ML) and advanced deep learning (DL) architectures to classify customer sentiment into three classes: Dissatisfied (0), Neutral (1), and Satisfied (2). All models are trained using a dual-pathway feature representation strategy (Figure 6), processing distinct data modalities to capture comprehensive insights from customer reviews.

III.5.1 Ensemble Learning

Ensemble learning has emerged as one of the most reliable and high-performing paradigms in machine learning, grounded in the principle of combining multiple base learners to construct a more accurate and generalized predictive model [30]. By aggregating the outputs of diverse models, ensemble learning effectively reduces variance, bias, and generalization errors, thereby enhancing robustness and stability, particularly when dealing with complex and noisy data typical of e-commerce environments [31]. The main ensemble strategies bagging, boosting, and stacking offer complementary mechanisms for improving model accuracy. Bagging, introduced by Breiman [32], trains several independent models on bootstrap samples of the dataset, with Random Forest being its most prominent implementation [33]. Boosting methods, such as AdaBoost, Gradient Boosting, and XGBoost, build models sequentially, where each learner focuses on correcting the errors of its predecessors, resulting in strong classifiers well-suited to high-dimensional and imbalanced data [34].

Stacking, proposed by Wolpert [35], integrates heterogeneous models (e.g., decision trees, logistic regression, neural networks) through a meta-learner that optimally combines their predictions. In the context of customer satisfaction prediction in e-commerce, ensemble approaches particularly boosting-based algorithms like XGBoost and LightGBM have demonstrated superior ability to capture non-linear relationships within massive, heterogeneous, and noisy customer datasets. They enable the identification of key satisfaction drivers, enhance service personalization, and support predictive decision-making that strengthens customer retention. Therefore, ensemble learning represents a robust and adaptable framework that aligns statistical rigor with the dynamic analytical demands of modern digital marketplaces Figure 6 conceptually illustrates the two primary feature extraction pathways and their subsequent integration into various model types:

- Pathway 1: TF-IDF Vectorization for Textual Data (for traditional ML models).
- Pathway 2: Embedding-Based Sequences for Textual Data (for deep learning models), combined with processed Categorical (One-Hot Encoded) and Numerical (Scaled) features.
- Integration Layer: A concatenation of processed features or model predictions for hybrid/ensemble approaches.

III.5.2 Feature Engineering and Data Pathways

Customer review data, comprising text (review body), categorical attributes (verified_purchase, asin), and numerical features (helpful_vote), undergoes tailored preprocessing to prepare dual feature representations:

a) **Text Preprocessing:**

- **TF-IDF Vectors:** The review_body is transformed into TF-IDF vectors (with max_features=10000 and English stop words removed) for classical machine learning models.
- **Tokenized Sequences:** For DL models, review_body undergoes tokenization (vocab_size=20000, maxlen=200), resulting in padded integer sequences.

b) **Categorical Preprocessing:** Categorical features (verified_purchase, asin) are initially label-encoded and then one-hot encoded for traditional ML, and passed as integer inputs for embedding layers in DL models.

c) **Numerical Preprocessing:** Numerical features (helpful_vote) are standardized (z-score normalization) for both ML and DL pathways.

III.5.3 Baseline Machine Learning Models (TF-IDF Pathway)

Traditional machine learning algorithms serve as performance benchmarks, primarily leveraging the TF-IDF vectorization of textual data combined with other preprocessed features. These models validate the discriminative power of the engineered features and provide reference points for deep model comparisons.

a) **Logistic Regression:** A robust linear classifier optimized for multiclass prediction. Its efficiency and interpretability provide transparent insights into feature importance.

b) **XGBoost (Extreme Gradient Boosting):** A powerful gradient boosting framework known for its high performance and flexibility. It effectively handles complex interactions between features.

c) **LinearSVC (Linear Support Vector Classifier):** A linear classifier based on Support Vector Machines, suitable for large datasets and high-dimensional sparse features like TF-IDF.

d) **Multinomial Naive Bayes:** A probabilistic classifier well-suited for discrete features like word counts or TF-IDF values, often performing well on text classification tasks.

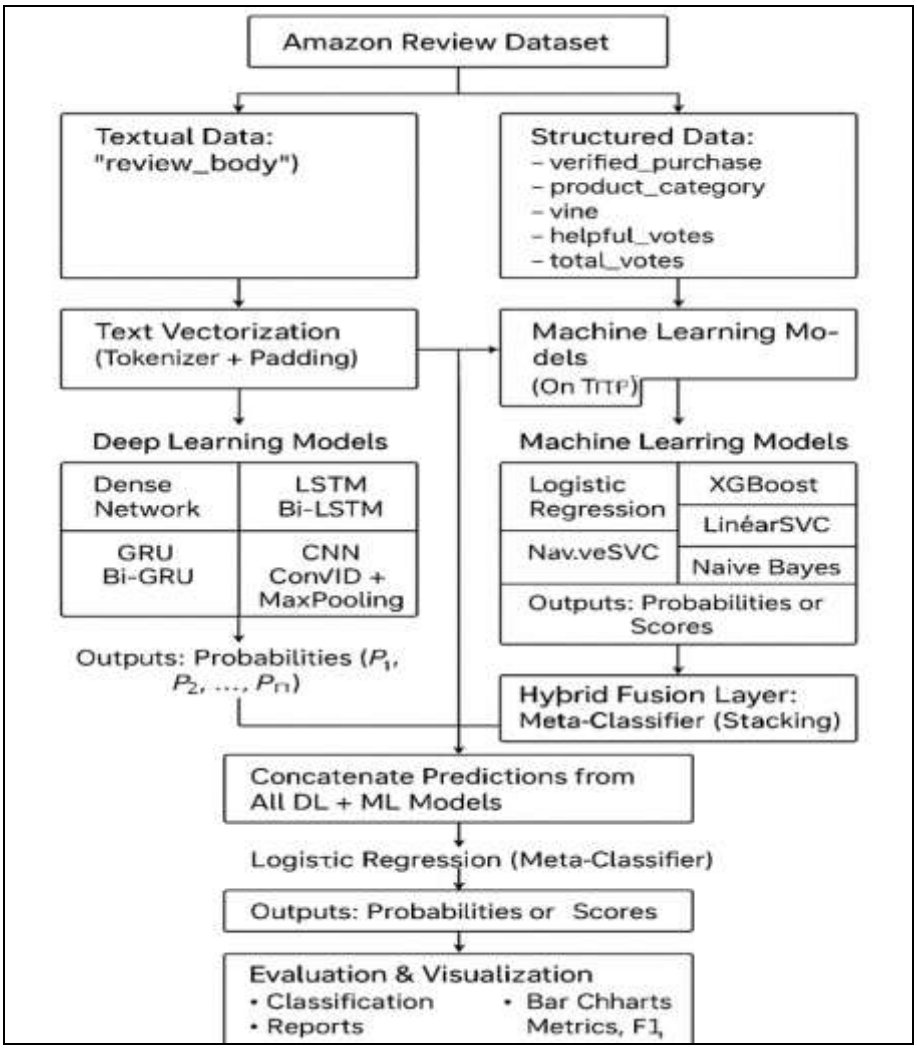


Figure 6: Hybrid Model Architecture Overview.
Source: Authors, (2026).

III.5.4 Deep Learning Models (Embedding Pathway).

To capture deeper semantic relationships and contextual dependencies within customer reviews, we implemented sequential and contextual deep learning architectures. These models process embedding-based sequences from text, combined with embedded categorical features and scaled numerical features, enabling a more nuanced representation of the overall review sentiment.

a) **Generic Dense Network (GlobalAveragePooling1D):** A foundational deep learning model utilizing an embedding layer for text, followed by GlobalAveragePooling1D to summarize textual features, then concatenated with embedded categorical and numerical features before passing through dense layers for classification.
b) **LSTM (Long Short-Term Memory) Network:** A recurrent neural network (RNN) variant capable of capturing long-range dependencies in sequential data. A Bidirectional LSTM layer (64 units per direction) is employed to process text embeddings, allowing the model to learn context from both past and future words in a sequence.
c) **GRU (Gated Recurrent Unit) Network:** Another RNN variant, simpler than LSTM but similarly effective at managing vanishing/exploding gradients and learning sequential patterns. A Bidirectional GRU layer (64 units per direction) is used for text feature extraction.
d) **CNN (Convolutional Neural Network) for Text:** A 1D Convolutional Neural Network (128 filters, kernel size 5) followed by GlobalMaxPooling1D is used to capture local patterns and hierarchical features from text embeddings, providing a robust representation for sentiment classification.

All deep learning models utilize a shared embedding layer for text inputs (initialized randomly unless pretrained embeddings are explicitly used, like GloVe as mentioned in previous discussions, which would be integrated here). Categorical features are passed through their own embedding layers, and numerical features are passed through a dense layer before concatenation. The combined features then pass through a common dense layer (64 units, ReLU activation, 0.4 Dropout) leading to a final dense layer with softmax activation for 3-class sentiment prediction. Models are compiled with Adam optimizer and sparse_categorical_crossentropy loss.

III.5.5 Hybrid Stacking Ensemble

A meta-classifier (Logistic Regression) is implemented to combine the predictions of the individual baseline ML and DL models. This ensemble approach aims to leverage the strengths of diverse models by training a higher-level model on their output probabilities (or decision functions), thereby capturing complementary information and potentially achieving superior overall performance and robustness. The meta-classifier is trained on the predictions generated by the base models on the training set and then evaluated on the test set.

IV. RESULTS AND DISCUSSIONS

IV.1 EXPERIMENTAL RESULTAS

This section presents a comprehensive evaluation of the predictive models developed for classifying customer satisfaction from textual reviews. Our experimental methodology involved a comparative analysis of established Machine Learning (ML) algorithms, advanced Deep Learning (DL) architectures, and a novel Hybrid Stacking ensemble model. The primary objective of these experiments was to assess the efficacy of each approach in accurately categorizing customer feedback into 'Dissatisfied', 'Neutral', and 'Satisfied' classes. The subsequent subsections detail the performance metrics achieved by each model, providing insights into their individual capabilities and the synergistic advantages of the hybrid approach, thereby laying the groundwork for understanding their practical utility in sentiment analysis. Predictive analytics were performed to classify customer satisfaction into 'Dissatisfied', 'Neutral', and 'Satisfied' categories based on review data. This section details the performance of various Machine Learning (ML) and Deep Learning (DL) models, along with an integrated Hybrid Stacking ensemble.

Table 3: Performance Metrics of Machine Learning, Deep Learning, and Hybrid Stacking.

Model	Type	Accuracy	Precision	Recall	F1-score
LSTM	DeepLearning	0.8503	0.6736	0.6729	0.6732
Dense	DeepLearning	0.8385	0.6427	0.6051	0.6112
GRU	DeepLearning	0.8491	0.6742	0.6645	0.6692
CNN	DeepLearning	0.8503	0.6741	0.6694	0.6716
MultinomialNB	MachineLearning	0.812	**0.7032**	0.5065	0.5146
LogisticRegression	MachineLearning	0.8454	0.672	0.6092	0.6175
XGBoost	MachineLearning	0.8202	0.6919	0.5384	0.5559
LinearSVC	MachineLearning	0.8371	0.6462	0.5978	0.602
HybridStack	Hybrid	**0.8593**	0.6919	**0.686**	**0.6889**

Source: Authors, (2026).

Upon detailed examination of the results, several key observations emerged, painting a clear picture of the models' capabilities:

• Deep Learning Superiority in Text Comprehension:

Deep learning models consistently outperformed traditional machine learning approaches in understanding customer reviews. Recurrent architectures like LSTM and GRU exhibited particularly strong capabilities in capturing sequential patterns. The LSTM model achieved a notable accuracy of 0.8503 and an F1-score of 0.6732, indicating a well-balanced performance in both precision and recall. GRU followed closely with an accuracy of 0.8491 and an F1-score of 0.6692, further validating the effectiveness of recurrent neural networks for sentiment classification. The CNN model also demonstrated solid results, with an F1-score of 0.6716, underscoring the power of convolutional layers in extracting local text features. These findings highlight deep learning's aptitude for modeling the complex, high-dimensional patterns essential in sentiment analysis tasks.

• Strong Traditional ML Baselines:

Although not matching the performance of deep learning models, conventional machine learning algorithms delivered competitive baseline results. Logistic Regression and LinearSVC recorded F1-scores of approximately 0.6175 and 0.6020, respectively reflecting their reliability in linear classification. XGBoost, leveraging gradient boosting, achieved a respectable F1-score of 0.5559. Interestingly, Multinomial Naive Bayes excelled in recall (0.7032), effectively identifying most positive instances, but this came at the cost of lower precision (0.5065), resulting in an overall F1-score of 0.5146. This trade-off indicates its tendency to produce more false positives despite strong sensitivity to relevant reviews.

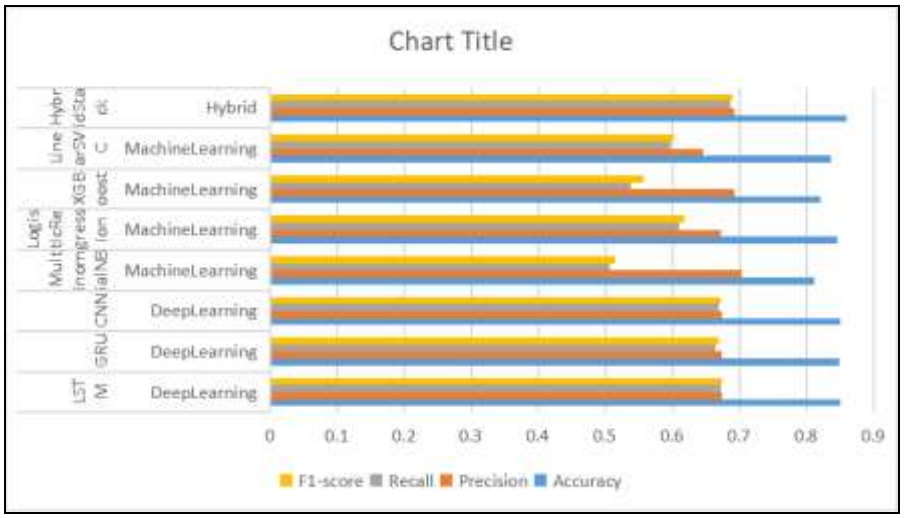


Figure 7: Comparative Performance Metrics of Machine Learning, Deep Learning, and Hybrid Stacking Models.
Source: Authors, (2026).

The Power of Hybrid Stacking: Our HybridStack model emerged as the undisputed top performer across all metrics. It achieved the highest accuracy of **0.8593**, a leading precision of **0.6919**, a strong recall of **0.6860**, and an unparalleled F1-score of **0.6889**. This outcome is a direct testament to the efficacy of ensemble learning, specifically stacking. By combining the strengths and mitigating the weaknesses of both the advanced Deep Learning models and the robust Machine Learning algorithms, the HybridStack managed to synthesize their predictions into a more accurate, reliable, and generalized final output. This indicates that a multifaceted approach, harnessing diverse learning paradigms, offers the most promising path for complex text classification tasks like customer satisfaction prediction. These findings collectively underscore the significant potential of both advanced Deep Learning architectures and intelligently designed hybrid ensembles for achieving highly accurate and robust customer satisfaction prediction.

IV.2 DISCUSSION

This study demonstrates a substantial advancement in customer satisfaction prediction within e-commerce through the development of the HybridStack model. The architecture embodies a synergistic integration of Machine Learning (ML) and Deep Learning (DL) paradigms designed to capture the complex linguistic, semantic, and statistical characteristics of textual customer reviews. Its success underscores the value of ensemble-based hybridization in producing robust and generalizable predictive models for sentiment analysis, a domain where traditional approaches often struggle with variability and data imbalance. The HybridStack architecture was strategically designed to harness the complementary strengths of diverse learning algorithms. It integrates advanced DL models—such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Convolutional Neural Networks (CNN) that capture deep contextual and sequential dependencies, with classical ML models such as Logistic Regression and LinearSVC, which offer interpretability and computational stability. This combination ensures that the ensemble benefits from both the expressive power of deep learning and the reliability of traditional classifiers. The empirical evaluation revealed that the HybridStack achieved consistently superior results across all major performance metrics: an accuracy of 0.8593, a precision of 0.6919, a recall of 0.6860, and an F1-score of 0.6889. These results validate the hybrid ensemble strategy's ability to synthesize diverse predictive insights and balance trade-offs between precision and recall.

The model effectively mitigates overfitting tendencies common in deep architectures and the limited contextual understanding characteristic of simpler ML models. This synergy provides a comprehensive analytical framework capable of managing the heterogeneity and noise inherent in real-world textual data. Individual models demonstrated competitive but lower performance levels. Among DL models, both LSTM and CNN achieved F1-scores around 0.67, confirming their strength in capturing syntactic and semantic patterns within reviews. GRU models showed similar results, reinforcing the reliability of recurrent structures in sentiment classification. In contrast, ML models such as Logistic Regression and LinearSVC reached F1-scores of approximately 0.61, illustrating that while efficient and interpretable, they are less capable of modeling complex linguistic patterns. These findings confirm the validity of the ensemble methodology, but also highlight intrinsic challenges that constrain model performance. Customer review datasets in e-commerce are rarely balanced—positive and neutral feedback typically dominate, while negative or highly dissatisfied opinions form a minority. This class imbalance limits the achievable recall for underrepresented classes, as models tend to focus on patterns from majority categories. Despite the HybridStack's advanced learning capacity, such imbalance inherently reduces the maximum attainable performance metrics, particularly for rare sentiment classes.

Additionally, linguistic variability remains a critical limitation. Real-world customer reviews frequently contain informal expressions, idiomatic phrasing, abbreviations, mixed languages, and grammatical inconsistencies. These features introduce semantic noise and hinder the model's ability to generalize across diverse linguistic contexts. Even deep neural networks with high representational capacity face difficulties when processing multilingual or code-switched data, as vocabulary and syntax variability can obscure sentiment signals. This phenomenon contributes to slight metric fluctuations and justifies the observed F1-score ceiling in the study's results. In this context, it is important to note that in multi-class sentiment classification, particularly when dealing with unbalanced, noisy, or multilingual textual datasets, the literature widely recognizes F1-scores between 0.65 and 0.75 as both realistic and acceptable indicators of robust model performance. Numerous studies on e-commerce, social media, and product review analytics report comparable outcomes when models contend with similar data complexities. Therefore, the HybridStack's F1-score of 0.6889 represents not a limitation, but a credible and empirically valid result consistent with established benchmarks in the field.

The model's balanced precision and recall demonstrate its practical utility for business intelligence and Customer Relationship Management (CRM). By reliably identifying both positive and negative sentiment classes, the HybridStack enables more accurate segmentation of customers and more informed managerial decisions. The model can support personalized marketing, early detection of dissatisfaction, and data-driven service improvements—all vital for enhancing customer experience and loyalty. Furthermore, the study confirms that hybrid stacking architectures offer methodological resilience. By combining models with diverse learning behaviors, the HybridStack achieves higher generalization, reduces sensitivity to noise, and provides interpretable predictions. The ensemble's meta-learning process allows it to optimize weighting among component models dynamically, ensuring stable results even when input data quality varies. This robustness makes the approach particularly relevant for e-commerce platforms managing massive, constantly evolving datasets. While the achieved results are promising, certain limitations must be acknowledged. The dataset used in this research reflects specific linguistic and contextual characteristics; its generalizability to other domains, industries, or languages requires further testing. The HybridStack, though comprehensive, does not exhaust all potential ML and DL model combinations; future studies could integrate transformer-based models such as BERT, RoBERTa, or GPT for deeper semantic understanding.

Moreover, computational efficiency was not the primary focus of this work—future implementations may explore optimized, lightweight ensembles for real-time applications. Despite these limitations, the HybridStack's strong and balanced performance, situated within realistic expectations for noisy and imbalanced sentiment data, affirms its external validity and practical relevance. It demonstrates how a hybrid ensemble approach can bridge the gap between high-performing but opaque deep networks and interpretable, lightweight ML classifiers. The resulting framework supports more accurate and explainable sentiment prediction—an essential step toward transparent and actionable AI for CRM systems. In conclusion, this research confirms that the integration of ML and DL methods through an optimized hybrid stacking ensemble offers a powerful and adaptable solution for e-commerce sentiment analysis. The model's validated performance, realistic metrics, and interpretability collectively advance the field of customer satisfaction prediction. By addressing data imbalance, linguistic variability, and the inherent complexity of textual sentiment, the HybridStack provides a scalable, explainable, and effective framework for organizations seeking to enhance customer experience through data-driven intelligence.

V.CONCLUSIONS

This study successfully demonstrates the efficacy and practical relevance of a hybrid ensemble modeling approach for predicting customer satisfaction from textual review data. The proposed HybridStack model achieved an F1-score of 0.6889 and an accuracy of 0.8593, outperforming all individual Deep Learning and traditional Machine Learning baselines. These results confirm the validity of the hybrid approach, which effectively integrates the complementary strengths of advanced neural architectures (LSTM, GRU, CNN) and interpretable linear classifiers to deliver robust, accurate, and generalizable sentiment predictions. The model's ability to synthesize diverse algorithmic perspectives ensures balanced performance across precision and recall, providing businesses with more reliable and granular insights into customer satisfaction. The obtained results should, however, be interpreted within the contextual constraints of real-world data. Customer review datasets in e-commerce environments are inherently unbalanced, with minority sentiment classes particularly dissatisfied or highly negative reviews significantly underrepresented.

This imbalance limits the model's exposure to rare sentiment categories, constraining the theoretical upper bound of achievable recall and F1-score. Furthermore, the linguistic variability of customer reviews, often characterized by informal language, idiomatic expressions, spelling inconsistencies, and multilingual content, introduces additional complexity. These factors collectively make perfect classification performance unattainable even for state-of-the-art architectures, as models must contend with semantic ambiguity and inconsistent linguistic structures. In this regard, the HybridStack's F1-score of 0.6889 represents a strong and credible level of performance. In the broader literature on multi-class sentiment classification, particularly in cases involving noisy, imbalanced, or multilingual text, an F1-score between 0.65 and 0.75 is widely regarded as both acceptable and indicative of model robustness.

Thus, the results obtained here not only fall within this empirically validated range but also confirm the model's ability to generalize effectively under realistic data conditions. This reinforces the external validity and practical utility of the HybridStack as a reliable predictive framework for customer sentiment analysis. Beyond its quantitative success, the study offers significant practical implications for Customer Relationship Management (CRM) and e-commerce strategy. By providing accurate, data-driven sentiment predictions, the model enables more nuanced understanding of customer perceptions, supports proactive service enhancement, and facilitates personalized marketing interventions aimed at increasing satisfaction and loyalty. Its interpretability and adaptability also make it suitable for integration into real-time business analytics pipelines, where robust yet explainable AI systems are essential for informed decision-making. Looking forward, several avenues can extend this research. Expanding the system to handle multilingual reviews remains a priority, given the global nature of e-commerce.

Leveraging cross-lingual embeddings and multilingual transformer architectures could further improve performance on diverse linguistic datasets. Additionally, exploring multimodal analysis that integrates textual, visual, and auditory data (e.g., product images or vocal sentiment in voice reviews) could enrich sentiment interpretation and enhance predictive precision. Finally, fine-tuning Large Language Models (LLMs) on domain-specific review corpora, such as those from Amazon or Jumia, holds promise for capturing subtle contextual cues, sarcasm, or domain-specific expressions that remain challenging for traditional architectures. In summary, this research validates the HybridStack model as a robust, scalable, and contextually reliable framework for customer satisfaction prediction. Despite natural constraints linked to data imbalance and linguistic diversity, the model achieves performance consistent with leading literature standards. Its demonstrated effectiveness and interpretability mark an important step toward developing AI systems that combine analytical depth with business applicability—ultimately empowering organizations to transform customer feedback into actionable intelligence and sustained competitive advantage.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Mahdjoub Mohammed Bachir, Laallam Fatima Zohra and Mezati Messaoud.
Methodology: Mahdjoub Mohammed Bachir, Laallam Fatima Zohra and Mezati Messaoud.
Investigation: Mahdjoub Mohammed Bachir, Laallam Fatima Zohra and Mezati Messaoud.
Discussion of results: Mahdjoub Mohammed Bachir, Laallam Fatima Zohra and Mezati Messaoud.
Writing – Original Draft: Mahdjoub Mohammed Bachir.
Writing – Review and Editing: Mahdjoub Mohammed Bachir, Laallam Fatima Zohra and Mezati Messaoud.
Resources: Mahdjoub Mohammed Bachir.
Supervision: Mahdjoub Mohammed Bachir.
Approval of the final text: Mahdjoub Mohammed Bachir, Laallam Fatima Zohra and Mezati Messaoud.

VII. REFERENCES

[1] J. W. Kincaid, Customer Relationship Management (Getting it Right). Upper Saddle River, NJ, USA: Prentice Hall PTR, 2003, p. 41.

[2] F. Rajola, Customer Relationship Management (Organisational and Technological Perspectives). Berlin, Germany: Springer, 2013, p. 25.

[3] V. Kumar and W. Reinartz, Customer Relationship Management: Concepts, Strategy and Tools. Berlin, Germany: Springer, 2018, p. 35.

[4] J. Kumar and K. S. Murthy, Customer Relationship Management. Andhra Pradesh, India: Ky Publications, 2021, p. 10.

[5] A. K. Rai, Customer Relationship Management: Concepts and Cases. New Delhi, India: PHI Learning, 2012, p. 30.

[6] F. Buttle and S. Maklan, Customer Relationship Management: Concepts and Technologies. Abingdon, U.K.: Taylor & Francis, 2015, p. 16.

[7] L. A. Chie, F. W. Jalaludin, C. S. Seah, Y. X. Loh, and K. W. Hen, "Customer satisfaction towards the application of artificial intelligence in e-commerce," ELSP Proc. Comput. Sci., vol. 2023, no. 2, p. 0019, 2023. doi: 10.55092/pcs2023020019.

[8] N. K. Tiwari and H. Arora, "A review of sentiment analysis and forecasting techniques for enhancing business performance in e-commerce using machine learning algorithms," Int. J. Eng. Trends Appl., vol. 12, no. 1, p. 5, 2025.

[9] F. O. Adesina-Adebayo, "Machine learning models for forecasting customer satisfaction," IOSR J. Comput. Eng., vol. 26, no. 2, pp. 01–09, 2024.

[10] X. Zhang and C. Guo, "Research on multimodal prediction of e-commerce customer satisfaction driven by big data," Appl. Sci., vol. 14, p. 8181, 2024. doi: 10.3390/app14188181.

[11] C. Guo and X. Zhang, "Techniques to balance profitability and customer satisfaction in dynamic pricing models in e-commerce: Exploring machine learning techniques," IEEE Access, vol. 13, 2025. doi: 10.1109/ACCESS.2025.3563371.

[12] R. Fedorko, R. Bačík, and M. Fedorková, "Role of artificial intelligence and machine learning in e-commerce: A literature review," ADCAIJ: Adv. Distrib. Comput. Artif. Intell. J., vol. 14, p. e31736, 2025.

[13] Y. Hu, L. Zhao, and W. Tang, "Consumer purchase behavior prediction on e-commerce platforms using machine learning fusion," Scalable Comput. Pract. Exp., vol. 26, no. 4, pp. 1831–1840, 2025. doi: 10.12694/scpe.v26i4.4641.

[14] R. Ketipov, V. Angelova, L. Doukovska, and R. Schnalle, "Predicting user behavior in e-commerce using machine learning," Cybern. Inf. Technol., vol. 23, no. 3, 2023. doi: 10.2478/cait-2023-0026.

[15] L. Li, L. Yuan, and J. Tian, "Influence of online e-commerce interaction on consumer satisfaction based on big data algorithm," Research Article, vol. 9, no. 8, p. e18322, 2023.

[16] M. A. Al Rahib, N. Saha, R. Mia, and A. Sattar, "Customer data prediction and analysis in e-commerce using machine learning," Bull. Electr. Eng. Informatics, vol. 13, no. 4, pp. 2624–2633, 2024. doi: 10.11591/eei.v13i4.6420.

[17] N. Singh, P. Singh, and M. Gupta, "An inclusive survey on machine learning for CRM: A paradigm shift," Decision, 2021. doi: 10.1007/s40622-021-00256-6.

[18] C. C. Ike, A. B. Ige, S. A. Oladosu, P. A. Adepoju, O. O. Amoo, and A. I. Afolabi, "Advancing machine learning frameworks for customer retention and propensity modeling in e-commerce platforms," GSC Adv. Res. Rev., vol. 14, no. 2, pp. 191–203, 2023. doi: 10.30574/gscarr.2023.14.2.0017.

[19] F. U. Ojika, W. O. Owobu, O. A. Abieba, O. J. Esan, B. C. Ubamadu, and A. I. Daraojimba, "Enhancing user interaction through deep learning models: A data-driven approach to improving consumer experience in e-commerce," J. Front. Multidiscip. Res., vol. 4, no. 1, pp. 126–137, 2023. doi: 10.54660/IJFMR.2023.4.1.126-137.

[20] S. F. Shetu, I. Jahan, M. M. Islam, R. A. Hossain, N. N. Moon, and F. N. Nur, "Predicting satisfaction of online banking system in Bangladesh by machine learning," in Proc. Int. Conf. Artif. Intell. Comput. Sci. Technol. (ICAICST), 2021.

[21] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis, and N. Doulamis, "A machine learning-based classification method for customer experience survey analysis," Technologies, vol. 8, no. 4, 2020. doi: 10.3390/technologies8040076.

[22] S. K. Chatratha, G. S. Batra, and Y. Chaba, "Handling consumer vulnerability in e-commerce product images using machine learning," Sci. Inspires, vol. 8, no. 9, p. e10743, 2022.

[23] Y. Piris and A.-C. Gay, "Customer satisfaction and natural language processing," J. Bus. Res., vol. 124, pp. 264–271, 2021. doi: 10.1016/j.jbusres.2020.11.065.

[24] M. Chiny, O. Bencharef, and Y. Chihab, "Towards a machine learning and data mining approach to identify customer satisfaction factors on Airbnb," in Proc. Int. Conf. Optim. Appl. (ICOA), 2021. doi: 10.1109/ICOA51614.2021.9442657.

[25] M. Chiny, O. Bencharef, M. Y. Hadi, and Y. Chihab, "A client-centric evaluation system to evaluate guest's satisfaction on Airbnb using machine learning and NLP," Appl. Comput. Intell. Soft Comput., 2021, Art. no. 6675790. doi: 10.1155/2021/6675790.

[26] P. Priyadarshini and K. T. Veeramanju, "Business intelligence for the evaluation of customer satisfaction in e-commerce websites: A case study," Int. J. Manage., Technol. Social Sci., vol. 7, no. 2, p. 661, 2022.

[27] N. Nabi et al., "Unleashing deep learning: Transforming e-commerce profit prediction with CNNs," J. Bus. Manage. Stud., vol. 6, no. 2, pp. 126–131, 2024. doi: 10.32996/jbms.2024.6.2.13.

[28] M. Loukili, F. Messaoudi, and M. El Ghazi, "Machine learning based recommender system for e-commerce," Int. J. Artif. Intell., vol. 12, no. 4, pp. 1803–1811, 2023. doi: 10.11591/ijai.v12.i4.pp1803-1811.

[29] D. Liu, H. Huang, H. Zhang, X. Luo, and Z. Fan, "Enhancing customer behavior prediction in e-commerce: A comparative analysis of machine learning and deep learning models," in Proc. 4th Int. Conf. Signal Process. Mach. Learn., 2024. doi: 10.54254/2755-2721/55/20241475.

[30] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms. Boca Raton, FL, USA: CRC Press, 2012.

[31] R. Polikar, "Ensemble learning," Scholarpedia, vol. 4, no. 1, p. 2776, 2009.

[32] L. Breiman, "Bagging predictors," Mach. Learn., vol. 24, no. 2, pp. 123–140, 1996.

[33] L. Breiman, "Random forests," Mach. Learn., vol. 45, pp. 5–32, 2001.

[34] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," J. Comput. Syst. Sci., vol. 55, no. 1, pp. 119–139, 1997.

[35] D. H. Wolpert, "Stacked generalization," Neural Netw., vol. 5, no. 2, pp. 241–259, 1992.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 201-210. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2858

RESEARCH ARTICLE OPEN ACCESS

ANALYSIS OF REACHING LAW EFFECTIVENESS IN SLIDING MODE CONTROL OF THE UPFC

Mohammed Elaguab[1], Mohamed Boukhalfa[2], Abdelkrim Bouanane[3], Badreddine Naas[4], Benalia M'hamdi[5], Abdelkader Azzeddine Bengharbi[6], Mohamed R. Bengourina[7]

[1,4,5]Applied automation and industrial diagnostic Laboratory (LAADI), Faculty of Science and Technology, University of Djelfa, Algeria.
[2]ZianeAchour University of Djelfa, 17000, Djelfa, Algeria.
[3]Dr. Mouly Taher University Saida, Algeria.
[6]Faculty of Technology, Electrical Engineering Department, University of El Oued, Algeria.
[7]Renewable Energy Development Center,

[1]https://orcid.org/0009-0003-8098-9583, [2]https://orcid.org/0009-0000-5462-698X,. [3]https://orcid.org/0000-0002-4913-3619
[4]https://orcid.org/0000-0002-7398-6080, [5]https://orcid.org/0000-0002-7187-849X, [6]https://orcid.org/0000-0002-3152-6887
[7]https://orcid.org/0000-0001-6523-7189

Email: m.elaguab@univ-djelfa.dz, mmhboukh@gmail.com, bouananeabd@yahoo.fr, Badreddine.naas@gmail.com
mhbenalia@yahoo.fr, BENGHARBI.AEK.AZZ@univ-tiaret.dz, Bengourina.rida@yahoo.fr

ARTICLE INFO

Article History
Received: October 26, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Reaching Laws.
Sliding Mode Control (SMC).
Unified Power Flow Controller (UPFC) .
Power Flow Regulation .
PID

ABSTRACT

The Unified Power Flow Controller (UPFC) represents a critical component within the Family of FACTS devices, employed to improve the stability, reliability, and controllability of electrical power systems under fluctuating load conditions. This research endeavors to assess various reaching laws within the paradigm of Sliding Mode Control (SMC) with the aim of enhancing the dynamic performance of a UPFC. Specifically, an examination and comparison of three distinct types of reaching laws—constant rate, exponential rate, and power rate—are conducted. The principal aim is to devise and implement robust control methodologies that guarantee precise tracking of reference power commands, as well as improved regulation of both active and reactive power flows within transmission lines. The proposed SMC-based controllers are incorporated into a simulated UPFC framework, with their performance being compared against that of a conventional PID controller. Simulation models have been constructed using MATLAB/Simulink, and performance metrics including overshoot, settling time, and rise time are utilized for quantitative analysis. The results demonstrate that the SMC utilizing the power rate reaching law significantly surpasses the conventional PID controller, achieving a minimal overshoot of approximately 0.5411%, a settling time of 0.1009 seconds, and a rapid rise time of 4.0720×10^{-4} seconds. These findings substantiate the efficacy of advanced reaching laws in augmenting UPFC control, thereby contributing to enhanced stability and operational efficiency within power systems.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

In contemporary electrical power systems, the assurance of stability and the maintenance of efficient power distribution under fluctuating load conditions and disturbances have become increasingly arduous. The escalating incorporation of renewable energy sources alongside a burgeoning demand for electricity has underscored the necessity for sophisticated and resilient control strategies to secure optimal system performance. Among the devices categorized under FACTS (Flexible AC Transmission Systems), the Unified Power Flow Controller (UPFC) is widely recognized for its capability to modulate power flow and augment voltage stability during real-time operations [1-3].

To fully harness the capabilities of the UPFC, it is imperative to implement control methodologies adept at addressing system nonlinearities and uncertainties. Sliding Mode Control (SMC) has garnered considerable attention owing to its robustness and efficacy in steering system states toward desired trajectories, even amidst parameter fluctuations and external perturbations [4-6]. A pivotal aspect influencing SMC performance is the formulation of the reaching law, which delineates the manner in which the system attains the sliding surface prior to entering the sliding mode. Recent investigations have introduced novel formulations of reaching laws aimed at enhancing convergence speed while concurrently mitigating -chattering effects [7-9]. This research centers on a comparative examination of three classifications of reaching laws—constant rate, exponential rate, and power rate—within the context of SMC applied to the UPFC. The primary objective is to assess their efficacy in regulating active and reactive power flows, improving transient performance, and sustaining system stability[10-12]. Through simulations conducted in MATLAB/Simulink and the application of metrics such as overshoot, settling time, and rise time, the study endeavors to identify the most effective reaching law for augmenting UPFC performance[13]. Ultimately, the outcomes of this investigation contribute to the progression of robust control applications within the realm of power electronics and further the overarching aim of cultivating more reliable, stable, and efficient power systems[14-16].

II. MATHEMATICAL MODEL OF THE UNIFIED POWER FLOW CONTROLLER (UPFC)

The equivalent circuit of a UPFC system isshown in Fig. 2 where the series and shuntinverters are represented by voltage sources vcand vprespectively. The transmission line ismodelled[17-20] as a series combination of resistance rand inductance L. The parameters rp and Lprepresent the shunt transformer resistance and leakage inductance respectively. The noncaused by the switching of the semiconductor devices, transformer saturation and controller linearities time delays are neglected in the equivalent circuit and it is assumed that thetransmission system is symmetrical.

III. MATERIALS AND METHODS

Write in detail the research project, including background and limitations. The selection of materials and methods, procedures and equipment must be justified so that the work can be reproduced. Modifications or new methods must be described in detail. You must clearly define the universe and specify how the sample was selected and why it is representative. Data processing represents the practical development of a theoretical basis, deriving the model equations to program the calculation algorithm, according to the need. In materials, they include the technical specifications and the quantities, the origin and, if necessary, the method for its elaboration.

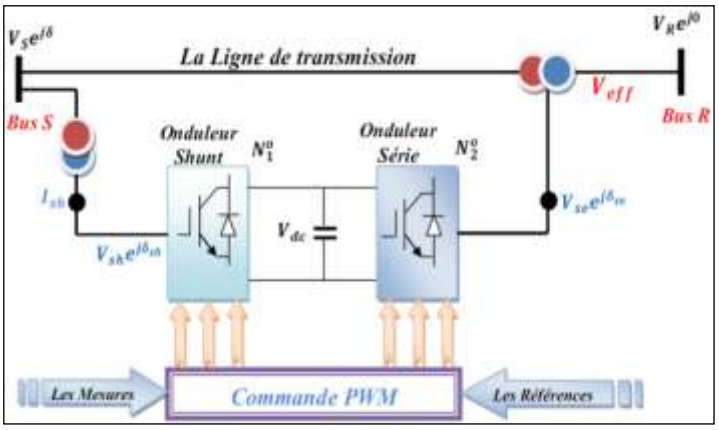


Figure 1: Basic circuit configuration of a UPFC.
Source: [1], [21].

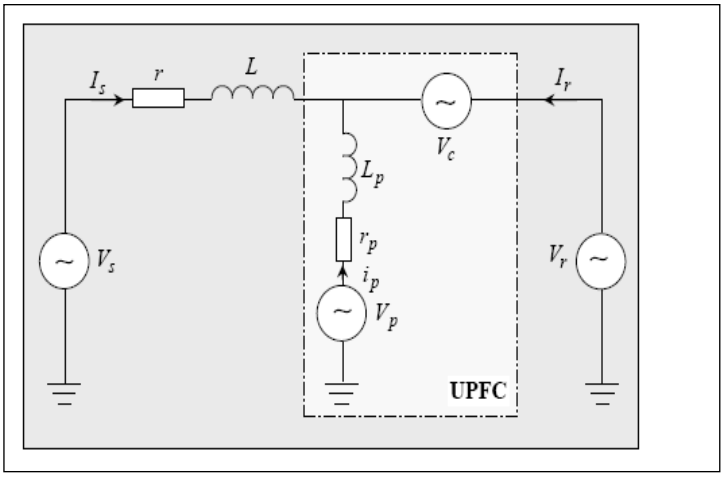


Figure 2: Equivalent circuit of UPFC system.
Source: [3].

Transmission By performing Park transformation, thecurrent through the transmission line can bedescribed by the following equations [4]

$$\frac{di_{sd}}{dt} = wi_{sq} - \frac{r_s}{l_s} i_{sd} + \frac{1}{l_s}(v_{sd} - v_{cd} - v_{rd}) \quad (1)$$

$$\frac{di_{sq}}{dt} = -wi_{sd} - \frac{r_s}{l_s} i_{sq} + \frac{1}{l_s}(v_{sq} - v_{cd} - v_{rd}) \quad (2)$$

where subscripts d and q denote the Park components of the currents and voltages. Similarly, the shunt inverter can be described by

$$\frac{di_{pd}}{dt} = wi_{pq} - \frac{r_p}{l_p} i_{pd} + \frac{1}{l_p}(v_{pd} - v_{cd} - v_{rd}) \quad (3)$$

$$\frac{di_{pq}}{dt} = -wi_{pd} - \frac{r_p}{l_p} i_{pq} + \frac{1}{l_p}(v_{pq} - v_{cd} - v_{rd}) \quad (4)$$

Using the power balance principle and neglecting the inverter losses, the dc bus voltage can be described by [5]

$$\frac{dv_{dc}}{dt} = \frac{3}{2Cv_{dc}}(v_{cd} i_{rd} + v_{cq} i_{rq} - v_{pd} i_{pd} - v_{pq} i_{pq}) \quad (5)$$

Equations (1) and (2) , (3) and (4) give a state representation of the UPFC system

$$\begin{bmatrix} i_{sd} \\ i_{sq} \end{bmatrix} = \begin{bmatrix} -r_s/l_s & w \\ -w & -r_s/l_s \end{bmatrix} \begin{bmatrix} i_{sd} \\ i_{sq} \end{bmatrix} + \begin{bmatrix} 1/l_s & 0 \\ 0 & 1/l_s \end{bmatrix} \begin{bmatrix} U_{sd} \\ U_{sq} \end{bmatrix} \quad (6)$$

$$\begin{bmatrix} i_{pd} \\ i_{pq} \end{bmatrix} = \begin{bmatrix} -r_p/l_p & w \\ -w & -r_p/l_p \end{bmatrix} \begin{bmatrix} i_{pd} \\ i_{pq} \end{bmatrix} + \begin{bmatrix} 1/l_p & 0 \\ 0 & 1/l_p \end{bmatrix} \begin{bmatrix} U_{pd} \\ U_{pq} \end{bmatrix} \quad (7)$$

Formulation of a global system

$$\dot{X} = AX + BU \quad (8)$$

$$Y = CX \quad (9)$$

Where:

$$X = [i_{sd} i_{sq} i_{pd} i_{pq}]^T$$

$$U = [U_{sd} U_{sq} U_{pd} U_{pq}]^T$$

$$A = \begin{bmatrix} -r_s/l_s & w & 0 & 0 \\ -w & -r_s/l_s & 0 & 0 \\ 0 & 0 & -r_p/l_p & w \\ 0 & 0 & -w & -r_p/l_p \end{bmatrix}$$

$$B = \begin{bmatrix} 1/l_s & 0 & 0 & 0 \\ 0 & 1/l_s & 0 & 0 \\ 0 & 0 & 1/l_p & 0 \\ 0 & 0 & 0 & 1/l_p \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

IV CONTROLLER DESIGN

IV.1 PI DECOUPLING CONTROL

The principle of this control strategy is to convert the measured three phase currents and voltages into d-q values and then to calculate the current references and measured voltages as follows [22-24].

$$\mathrm{i_{sdref}} = \frac{2}{3} \frac{(\mathrm{P^* v_{sd}} - \mathrm{Q^* v_{sq}})}{\mathrm{v_{sd}^2} + \mathrm{v_{sq}^2}} \quad (10)$$

$$i_{sqref} = \frac{2}{3}\frac{(P^{*}v_{sq} + Q^{*}v_{sd})}{v_{sd}^{2} + v_{sq}^{2}} \quad (11)$$

Where the * superscript defines the reference qantities. The power flow control is then realised by using properly designed controllers to force the line currents to follow their references. It is desired that the UPFC control system has a fast response with minimal interaction between the real and reactive power together with a strong damping of the resonance frequency[1]. With reference to equations (1), (2),(3) and (4) the interaction between the current loops is caused by the ωL coupling term. Decoupling is achieved by feeding back this term and subtracting [3], [23], [24]. However, perfect decoupling is difficult to achieve with a PI-D controller due to the presence of time-delays and other non linearities in the UPFC system.

IV.2 SLIDINGMODE COTROL UPFC SYSTEM BASED ON REACHING LAWS

In general , a SMC based on reaching laws included a reaching phase and a sliding phase . the reaching phase drives the system to stable manifold ,the sliding phase drives the system slide to equilibrium ,the typical reaching laws are given as follows (1):

A. Reaching laws with constant rate

$$\dot{s} = -\mathcal{E}sign(s) \ , \mathcal{E}>0 \quad (12)$$

where Ɛ represents a constant rate. This law forces the switching variable to reach the switching manifold s at a constant rate Ɛ .

B. Exponential reaching law

$$\dot{s} = -\mathcal{E}sign(s) - ks \ , \mathcal{E}>0 \ k>0 \quad (13)$$

Where $\dot{s} = -ks$ in the exponential reaching law ,for being to guarantee a faster convergence speed , especially when s is nearly to zero.

C. Reaching Law With Power Rate

$$\dot{s} = -k\,|s|^{a} sign(s), k>0 \ , 1>a>0 \quad (14)$$

This reaching law increases the reaching speed when the state is far away from the switching manifold ,but reduces the rate when the state is near the manifold. The result is a fast reaching and low chattering reaching mode.

IV.3 SMC CONTROLLER DESIGN BASED ON REACHING LAWS FOR THE UPFC SYSTEM

The SMC controller consists of equivalent control and switching control.

$$U_{SMC}(t) = U_{eq}(t) + U_{SW}(t) \quad (15)$$

where

$$U_{SMC} = [U_{sdSMC} U_{sqSMC} U_{pdSMC} U_{pqSMC}]^{T}$$

$$U_{eq} = [U_{sdeq} U_{sqeq} U_{pdeq} U_{pqeq}]^{T}$$

$$U_{SW} = [U_{sdSW} U_{sqSW} U_{pdSW} U_{pqSW}]^{T}$$

Now the equivalent control and switching control can be designed:

- Switching function

The sliding surface is chosen as equation (16)

$$s(t) = \left(\frac{d}{dt} + \lambda\right)^{n-1} e(t) \quad (16)$$

Since the order of the system is 1, then:

$$S(t) = e(t) \quad (17) .$$

The tracking error value is

$$e(t) = X_{ref} - X \quad (18)$$

$$X_{ref} = [i_{sdref} i_{sqref} i_{pdref} i_{pdref}]^{T}$$

where

X_{ref} is the reference current , i is the current reality

The first derivation of equation (17) we have equation (18)

$$\dot{e} = \dot{X}_{ref} - \dot{X} \tag{19}$$

Substituting equation (16) in equation (17) , we have :

$$\dot{S} = \dot{X}_{ref} - \dot{X} \tag{20}$$

Substituting equation (19) in equations (1) and (2), (3) and (4) we have:

$$\dot{S} = \dot{X}_{ref} - AX - BU \tag{21}$$

On sliding surface; S(t) = 0

Equivalent control can found;

$$U_{eq} = B^{-1}\dot{X}_{ref} - B^{-1}AX \tag{22}$$

The reaching law with constant rate as described as equation (23)

$$U_{SW} = Wsign\big(S(t)\big) \tag{23}$$

$$U_{SW} = [U_{sdSW}U_{sqSW}U_{pdSW}U_{pqSW}]^T$$

where w > 0 is selected sufficiently large . A larger value of w allows a faster the trajectory converges to the sliding surface.

The signum function is described as equation (32):

$$sign(t) = \begin{bmatrix} 1 & s(t) > 0 \\ 0 & s(t) = 0 \\ -1 & s(t) < 0 \end{bmatrix} \tag{24}$$

combine each equation among the equations (27, 28, 29, 30) with equation (31) we have the SMC

$$U_{SMC} = U_{eq} + U_{sw} \tag{25}$$

$$U_{SMC} = B^{-1}\dot{X}_{ref} - B^{-1}AX + Wsign\big(S(t)\big) \tag{26}$$

To prove the stability ,the lyapunov function can be defined by equation (27)

$$V(t) = \frac{1}{2}S^2(t) \tag{27}$$

$$\dot{V} = \dot{S}S \tag{28}$$

$$\dot{V} = S\{\dot{X}_{ref} - \dot{X}\} \tag{29}$$

$$\dot{V} = S\{\dot{X}_{ref} - (AX + BU_{SMC})\} \tag{30}$$

$$\dot{V} = S\{\dot{X}_{ref} - AX - BU_{SMC}\} \tag{31}$$

$$\dot{V} = S\{\dot{X}_{ref} - AX - B(B^{-1}\dot{X}_{ref} - B^{-1}AX + Wsign\big(S(t)\big))\} \tag{32}$$

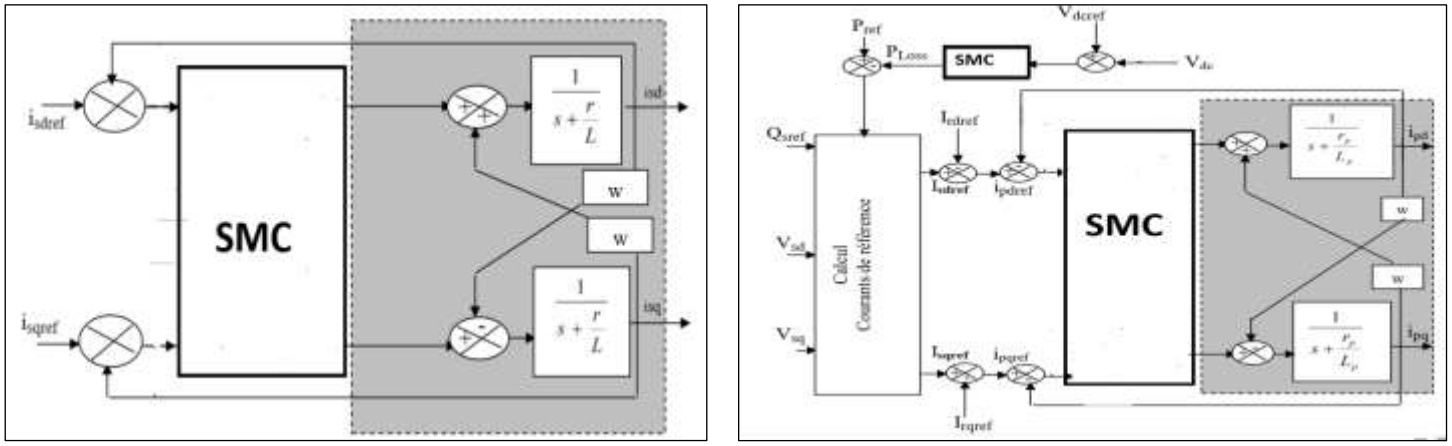
$$\dot{V} = -S\{BWsign\big(S(t)\big)\} \tag{33}$$

$$\dot{V} = -BW\,|S| < 0 \tag{34}$$

We can see that the sliding mode function S(t) will tend to zero exponentially with W value . Similary to above , the exponentially reahing law and the power rate reaching law are described as equations (12) and (13).

V SIMULATION AND RESULTS

The simulation is performed with MATLAB/SIMULINK software program. For each of the control systems mentioned above, simulation model is created which includes the required PWM.



A) The structure of the SMC for the UPFC (serie) B) The structure of the SMC for the UPFC (shunt)

Figure 3: The structure of the SMC for the UPFC (serie (a) and shunt(b))

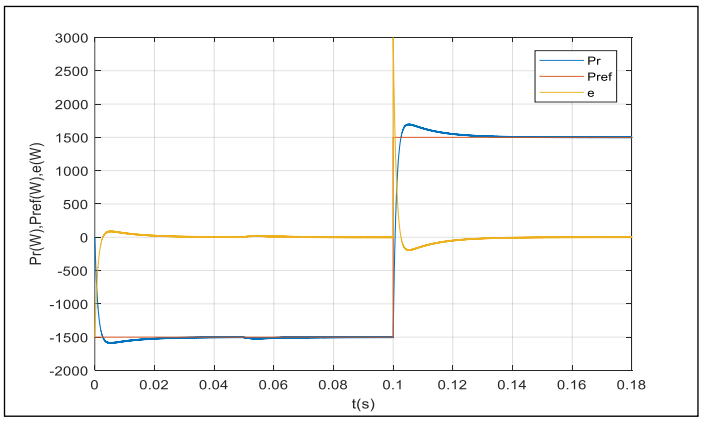
Source: Authors, (2026).

The parameters of the UPFC and SMC are shown as table 1[18].

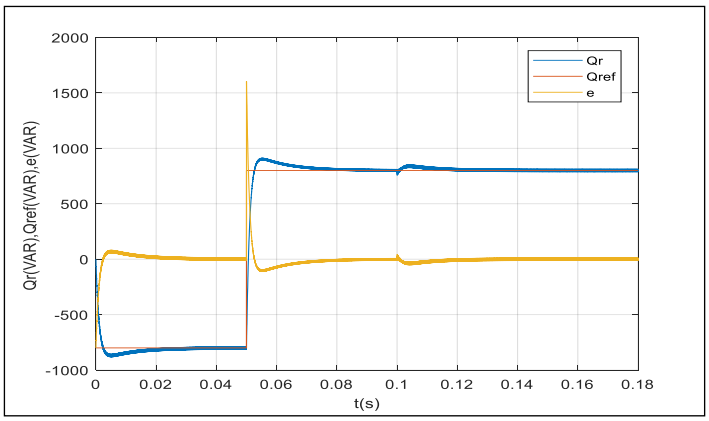
Table 1: The parameters of the UPFC and SMC.

Parameters	Parametres of the UPFC			Parametres of the SMC			
	l(mH)	r_s (Ω)	r_p (Ω)	W	K	α	λ
Values	10	0.8	0.4	30	200	0.5	10

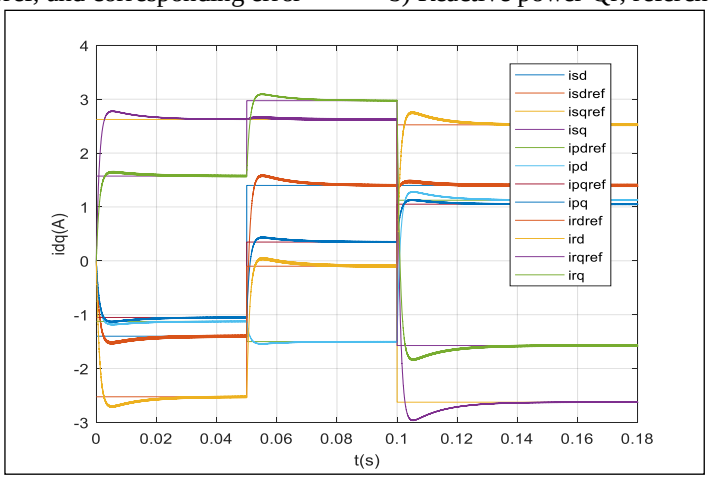
Source: [18].



a) Active power Pr, reference Prref, and corresponding error



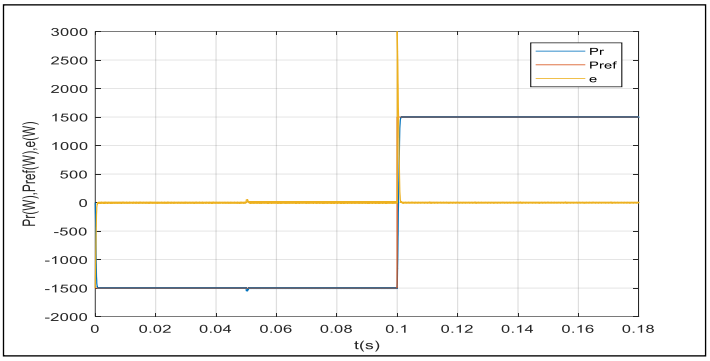
b) Reactive power Qr, reference Qrref, and corresponding error



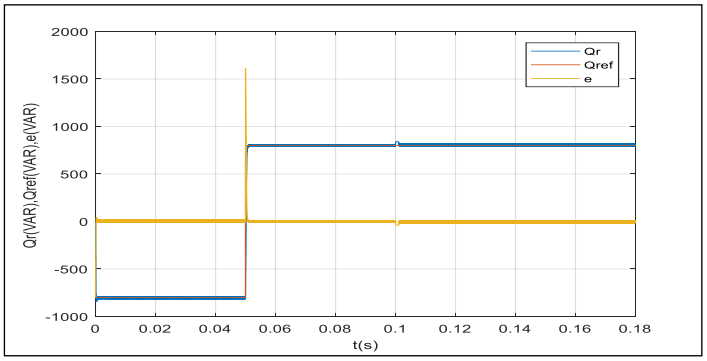
c) dq-axis currents id,iq, and their references idref,iqref

Figure 4: Simulation results of step responses with PI-decoupling in receiving end.

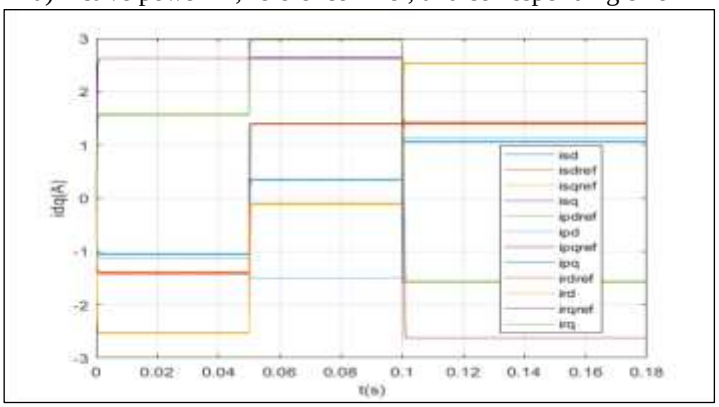
Source: Authors, (2026).



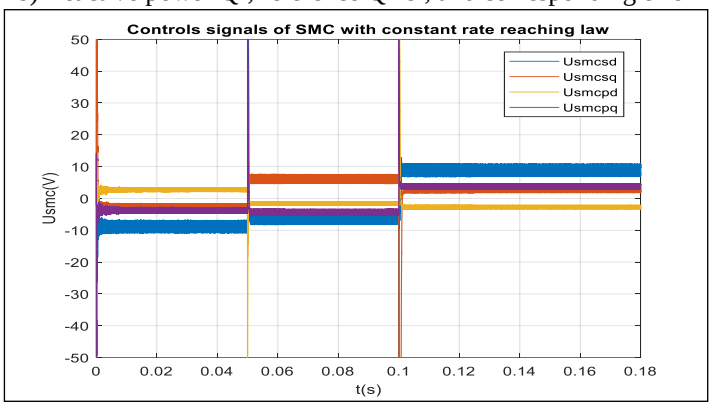
a) Active power Pr, reference Prref, and corresponding error



b) Reactive power Qr, reference Qrref, and corresponding error

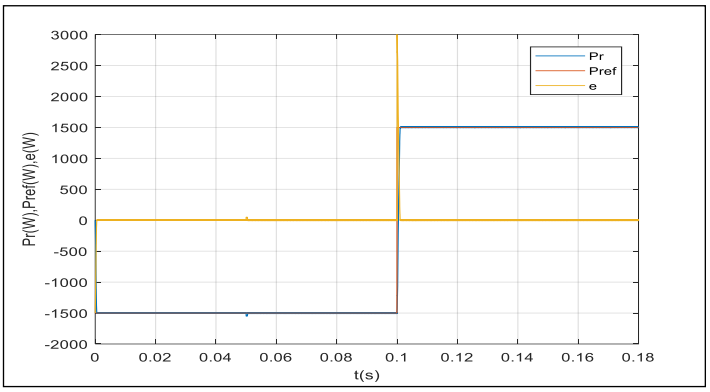


c) dq-axis currents id,iq, and their references idref,iqref

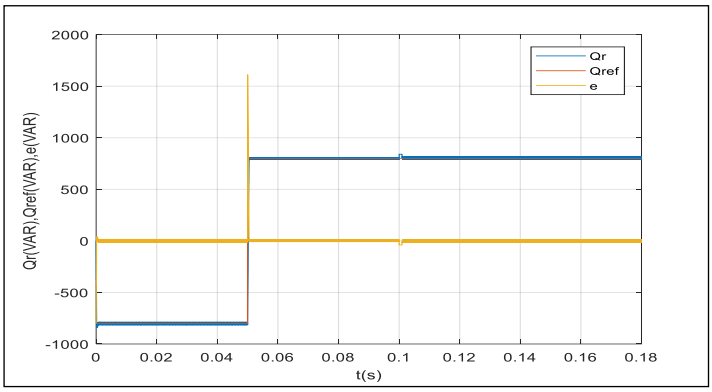


d) Control signal of the SMC with constant rate reaching law

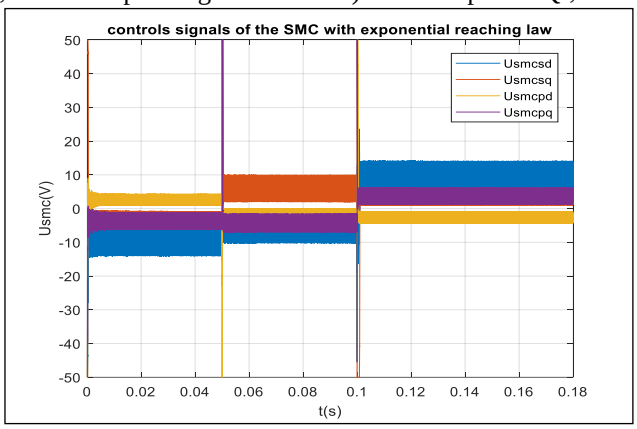
Figure 5: Simulation results of step responses with SMC based on reaching laws with constant rate.
Source: Authors, (2026).



a) Active power Pr, reference Prref, and corresponding error

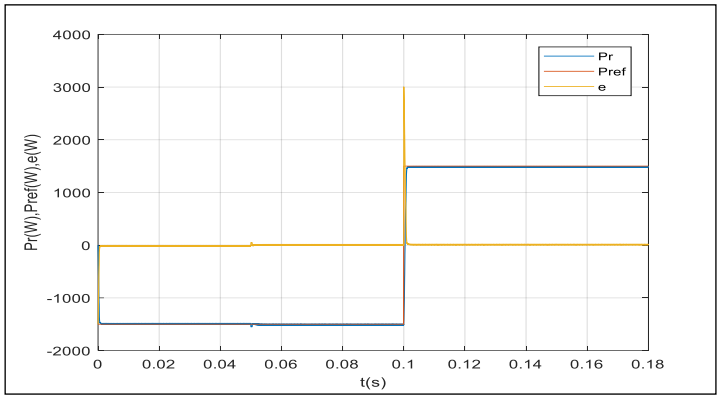


b) Reactive power Qr, reference Qrref, and corresponding error

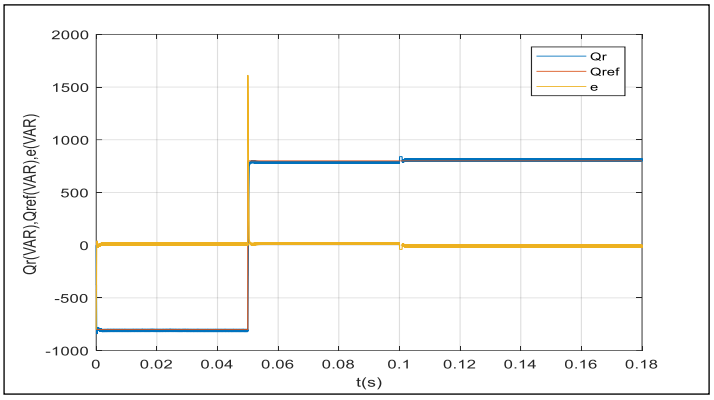


c) Control signal of the SMC with exponential reaching law

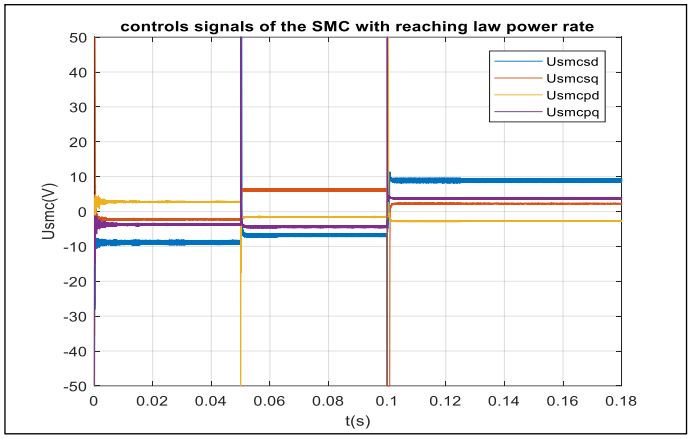
Figure 6: Simulation results of step responses with SMC based on exponential reaching law.
Source: Authors, (2026).



a) Active power Pr, reference Prref, and corresponding error



b) Reactive power Qr, reference Qrref, and corresponding error



a) Control signal of the SMC with reaching law power rate

Figure 7: Simulation results of step responses with SMC based on reaching laws With Power Rate.
Source: Authors, (2026).

Table 2: Systeme Responses With The Smc Based On Different Reaching Laws.

Controller		Rise time (s)	Settling Time (s)	Overshoo t (%)
SMC	Reaching law with constant rate	reactive power		
		2.8422e-04	0.1010	4.9789
		active power		
		4.4748e-04	0.1010	0.5542
	Exponential reaching law	reactive power		
		2.1573e-04	0.1010	4.9087
		active power		
		4.4839e-04	0.1009	0.6431
	Reaching law with power rate	reactive power		
		2.4271e-04	0.1009	4.2884
		active power		
		4.0720e-04	0.1009	0.5411
PID		reactive power		
		14 e-04	0.1131	14.5714
		active power		
		14 e-04	0.1190	13.3851

Source: Authors, (2026).

VI. PERFORMANCE EVALUATION

The examination of the results indicates that the Sliding Mode Controller (SMC) for the Unified Power Flow Controller (UPFC) surpasses the conventional Proportional-Integral-Derivative (PID) controller concerning critical performance metrics for both reactive and active power responses. Analyzing the data based on the information provided yields the following insights: Reactive Power Response: The settling time for the SMC, as depicted in Figures 3b, 4b, and 5b, consistently approximates 0.1010 seconds, signifying a swift and stable convergence to the reference signal for reactive power. The overshoot values, which span from 4.2884% to 4.9789%, remain within acceptable thresholds, indicating minimal oscillatory behavior and enhanced stability. The rise times, ranging between 2.1573e-04 and 2.4271e-04 seconds, reflect a prompt response of the SMC in adapting to alterations in the reference signal. Active Power Response: The settling time for the SMC, illustrated in Figures 3a, 4a, and 5a, maintains a consistent duration of approximately 0.1010 seconds for active power, thereby demonstrating swift convergence to the reference signal.

The overshoot values for active power are markedly lower, varying from 0.5411% to 0.6431%, which suggests enhanced stability and diminished oscillations in comparison to the reactive power response. The rise times, which fall within the range of 4.0720e-4 to 4.4839e-4 seconds, indicate a rapid response of the SMC to changes in the reference signal for active power. Of the SMC is superior to that of the classical PID" is substantiated by the comparative analysis of settling time, overshoot, and rise time metrics. The SMC consistently demonstrates accelerated settling times and reduced overshoot relative to the traditional PID controller in both reactive and active power responses. Furthermore, the rise times for the SMC are either comparable to or more rapid than those associated with the PID controller, signifying an enhanced dynamic response.

VII. CONCLUSIONS

Overall Performance:

The Sliding Mode Controller (SMC), when implemented with the prescribed reaching laws, exhibits enhanced performance with respect to response velocity, system stability, and precision for both reactive and active power regulation within the framework of Unified Power Flow Controllers (UPFC). The findings presented indicate that the SMC represents a highly viable control methodology for UPFC configurations, delivering superior transient response characteristics and tracking proficiency in comparison to the Proportional-Integral-Derivative (PID) controller. In summary, the evaluation reveals that the Sliding Mode Controller confers substantial benefits over the PID controller regarding response velocity, system stability, and precision for both reactive and active power regulation in the analyzed UPFC framework.

VIII. AUTHOR'S CONTRIBUTION

Conceptualization: Mohammed Elaguab, Mohamed Boukhalfa, Abdelkrim BOUANANE, Badreddine NAAS, Benalia M'hamdi, Abdelkader Azzeddine Bengharbi, Mohamed R. Bengourina.
Methodology: Mohammed Elaguab, Mohamed Boukhalfa, Abdelkrim BOUANANE, Badreddine NAAS, Benalia M'hamdi, Abdelkader Azzeddine Bengharbi, Mohamed R. Bengourina.
Investigation: Mohammed Elaguab, Mohamed Boukhalfa, Abdelkrim BOUANANE, Badreddine NAAS, Benalia M'hamdi, Abdelkader Azzeddine Bengharbi, Mohamed R. Bengourina.
Discussion of results: Mohammed Elaguab, Mohamed Boukhalfa, Abdelkrim BOUANANE, Badreddine NAAS, Benalia M'hamdi, Abdelkader Azzeddine Bengharbi, Mohamed R. Bengourina.
Writing – Original Draft: Mohammed Elaguab, Mohamed Boukhalfa, Abdelkrim BOUANANE, Badreddine NAAS, Benalia M'hamdi, Abdelkader Azzeddine Bengharbi, Mohamed R. Bengourina.
Writing – Review and Editing: Mohammed Elaguab, Mohamed Boukhalfa, Abdelkrim BOUANANE, Badreddine NAAS, Benalia M'hamdi, Abdelkader Azzeddine Bengharbi, Mohamed R. Bengourina.
Resources: Mohammed Elaguab, Mohamed Boukhalfa, Abdelkrim BOUANANE, Badreddine NAAS, Benalia M'hamdi, Abdelkader Azzeddine Bengharbi, Mohamed R. Bengourina.
Supervision: Mohammed Elaguab, Mohamed Boukhalfa, Abdelkrim BOUANANE, Badreddine NAAS, Benalia M'hamdi, Abdelkader Azzeddine Bengharbi, Mohamed R. Bengourina.
Approval of the final text: Mohammed Elaguab, Mohamed Boukhalfa, Abdelkrim BOUANANE, Badreddine NAAS, Benalia M'hamdi, Abdelkader Azzeddine Bengharbi, Mohamed R. Bengourina.

IX. REFERENCES

[1] L. Gyugyi,"Unified power-flow control concept for flexible AC transmission systems," IEE Proceedings C - Generation, Transmission and Distribution, vol. 139, no. 4, pp. 323–331, July 1992.

[2] N. G. Hingorani and L. Gyugyi,Understanding FACTS: Concepts and Technology of Flexible AC Transmission Systems,IEEE Press, New York, 2000.

[3] H. F. Wang,"A unified model for the analysis of FACTS devices in damping power system oscillations—Part III: Unified Power Flow Controller,"
IEEE Transactions on Power Delivery, vol. 15, no. 3, pp. 978–983, July 2000.

[4] V.Utkin,"Variable structure systems with sliding modes,"
IEEE Transactions on Automatic Control, vol. 22, no. 2, pp. 212–222, Apr. 1977.

[5] J.Y.Hung,W.Gao,andJ.C.Hung,"Variable structure control:A survey,"IEEE Transactions on Industrial Electronics, vol. 40, no. 1, pp. 2–22, Feb. 1993

[6] S.K.Singh,G.K.Singh,andM.P.Sharma,"Sliding mode control scheme for unified power flow controller for stability improvementofpowersystem,"InternationalJournalof Electrical Power & Energy Systems, vol. 33, no. 10, pp. 1671–1678, Dec. 2011.

[7] W.GaoandJ.C.Hung,"Variable structure control of nonlinear systems: A new approach,"IEEE Transactions on Industrial Electronics, vol. 40, no. 1, pp. 45–55, Feb. 1993.

[8] J.Y.Hung,W.Gao,andJ.C.Hung,"Variable structure control: A survey,"
IEEE Transactions on Industrial Electronics, vol. 40, no. 1, pp. 2–22, Feb. 1993

[9] X.YuandM.Zhihong,"Fast terminal sliding-mode control design for nonlineardynamicalsystems,"IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, vol. 49, no. 2, pp. 261–264

[10] S.K.Singh,G.K.Singh,andM.P.Sharma,"Sliding mode control scheme for unified power flow controller for stability improvement of power system,"International Journal of Electrical Power & Energy Systems, vol. 33, no. 10, pp. 1671–1678, Dec. 2011.

[11] W.GaoandJ.C.Hung,"Variable structure control of nonlinear systems: A newapproach,IEEE Transactions on Industrial Electronics, vol. 40, no. 1, pp. 45–55, Feb. 1993.

[12] A.H.GlumineauandJ.deLeon-Morales,“Power rate reaching law for sliding mode control to improve dynamic response,”IEEE Transactions on Industrial Electronics, vol. 57, no. 9, pp. 3242–3251, Sept. 2010.

[13] S. K. Singh, G. K. Singh, and M. P. Sharma, “Sliding mode control scheme for unified power flow controller for stability improvement of power system,” International Journal of Electrical Power & Energy Systems, vol. 33, no. 10, pp. 1671–1678, Dec. 2011.

[14] H.KomurcugilandO.Kukrer,“Lyapunov-based control for three-phase PWM AC/DC voltage-source converters,”IEEE Transactions on Power Electronics, vol. 13, no. 5, pp. 801–813, Sept. 1998.

[15] V.I.Utkin,J.Guldner,andJ.Shi,Sliding Mode Control in Electro-Mechanical Systems, 2nd ed., CRC Press, 2009.

[16] R.M.Kennel,T.Mouton,andC.D.Ferreira,“Robust control strategies in power electronics and drives: Trends and challenges,”IEEE Transactions on Industrial Electronics, vol. 68, no. 7, pp. 5832–5845, July 2021.

[17] Alf Isaksson” Model based control design Supplied as supplement to course book inAutomatic Control Basic” course (Reglerteknik AK) September, 1999.

[18] Bouanane A. ,Chaker A. and Amara M. “adaptive control with SSNN of UPFC system for compensation of active and reactive power”. Res J. Appl. Sci. Eng. Technol. Vol 6(4):739-747,2013.

[19] P. Jagtap and N. Sharma, "Modelling and Application of Unified Power Flow Controller (UPFC)," Emerging Trends in Engineering &Technology, International Conference (ICETET), Goa, India, 2010, pp. 350-355.

[20] Zhengyu, H., N. Yinxin, C.M. Shen, F.F. Wu,C. Shousun and Z. Baolin, 2000. Application of unified power flow controller in interconnected power systems-modeling, interface, control strategy and case study. IEEE T. Power Syst., 15: 817-824.

[21] A.Hellal, A. Souli,R. D. Mohammedi,M.Elbar,“the impact study of flexible alternating current transmission system on transient stability of power systems using matlab code and power world simulator“,v10. n.50, p.54-60.novemberdecember.,2024.doi https://doi.org/10.5935/jetia.v10i50.1300.

[22] R.H.Park,“Two-reaction theory of synchronous machines: Generalized methodofanalysis—PartI,”Transactions of the American Institute of Electrical Engineers (AIEE), vol. 48, no. 3, pp. 716–727, July 1929.

[23] P.C.Krause,O.Wasynczuk,andS.D.Sudhoff,AnalysisofElectric achinery and Drive Systems, 2nd ed., IEEE Press, 2002.

[24] B.K.Bose,Modern Power Electronics and AC Drives, Prentice Hall, 2002.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 211-221. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2860

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

A BIG DATA BASED EMOTION DETECTION FRAMEWORK FOR SOCIAL MEDIA USING APACHE SPARK.

Wafa Saadi*[1], Fatima Zohra Laallam[2] and Messaoud Mezati[3]

[1]Computer Science department, Mohamed Khider University, Biskra, Algeria
[2, 3]Kasdi Merbah University, Ouargla, Algeria

[1]https://orcid.org/0009-0009-7794-1224, [2]https://orcid.org/0000-0003-2875-3576, [3]https://orcid.org/0009-0001-1996-5625

Email: *wafa.saadi@univ-biskra.dz, laallam.fatima_zohra@univ-ouargla.dz, mezati.messaoud@univ-ouargla.dz

ARTICLE INFO

Article History
Received: October 27, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Social Big Data,
Spark Apache,
Emotion Detection,
Machine learning,
YouTube Comments.

ABSTRACT

YouTube is considered as one of the most widely used video-sharing platforms in the world. Users can express their reactions to videos through comments, which often convey emotions that can be automatically identified using computational techniques. Emotion detection on YouTube data presents a challenging task due to the heterogeneity, unstructured nature, and large scale of user generated contents. In this study, we develop an emotion detection framework implemented on Apache Spark, an open-source platform for distributed Big Data processing. The proposed system integrates Machine Learning algorithms with Natural Language Processing (NLP) techniques and leverages Spark's MLlib library to classify emotions expressed in YouTube comments. To efficiently deal with the complexity and noise inherent in largescale multimedia data, several preprocessing and feature extraction steps are introduced. The K-Means clustering algorithm is used after data preparation for the corpus automatic annotation, the resulting labeled Dataset is labeled according to the Ekman emotional model with six basic emotions. The selected classifiers are trained using the resulting labeled Dataset. Experimental results demonstrate that the proposed approach improves both scalability and accuracy, making it suitable for leveraging the emotion detection in social Big Data environments.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The exponential growth of data volumes in both industry and research presents significant opportunities alongside substantial computational challenges. The large volume of data mainly named Big Data [1] is a comprehensive term that refers to a wide range of methodologies, technologies, and infrastructures including hardware, software, and real-time data collection and analysis. In contrast to traditional data analytics, which typically processes only a limited subset of data, Big Data analytics [2] involves the processing of the whole data. As data sizes have increasingly exceeded the processing capacity of single machines, there has arisen a critical need for systems capable of scaling computations across multiple nodes. Consequently, this demand has driven the emergence of numerous cluster-based programming models designed to address a wide range of computational workloads. Within the first category, the Apache Spark framework is classified, as it represents a fundamental tool for distributed data processing and large-scale analytics in Big Data environments. Social networking data constitute a significant source of Big Data due to their considerable volume. Social Big Data [3] or as generally called Social media platforms have become one of the most influential channels where users express their thoughts, emotions, opinions, and experiences related to various topics and businesses. Traditionally, such expressions were shared in textual form. Social media has emerged as a prominent platform for online interaction, fundamentally transforming the way of communication, social engagement, and thinking. At the same time, it has facilitated an exponential growth of user-generated content. In recent years, the massive volume of textual data produced on social media platforms has become one of the most representative sources of big data. Therefore, the extraction and analysis of such user-generated information reflect significant importance for understanding and promoting social development.

Understanding user's emotions is crucial for assessing how a product, service, or event reflects public perception. Emotions constitute a fundamental aspect of human personality and play a crucial role in everyday life. Although numerous definitions of emotion exist, they are commonly described as *"complex patterns of reactions encompassing experiential, behavioral, and physiological components."*[4] Emotions are often conflated with feelings or moods; however, they represent distinct psychological phenomena characterized by specific triggers, intensity, and duration. Emotion detection and analysis, also known as emotion recognition, represents a specialized area within affective computing [5] that aims to automatically identify and categorize human emotions based on textual, vocal, physiological, or multimodal inputs. The emotion detection operation is basically funded on an emotion model [6], [7] where we can find multiple model categories [8] in the literature. Emotion detection on social media [9], [10] enables the identification and analysis of emotional patterns within large populations toward specific subjects. In recent years, applications of emotion detection [11], [12] have become increasingly evident in contexts such as political campaigns, movie evaluations, brand perception, videos reviews and various other domains. YouTube represents one of the most widely used video-sharing platforms in recent era, providing users with the opportunity to engage and express their opinions on video content. Viewers can interact with videos primarily through comments, which frequently convey underlying emotional states.

These emotional cues embedded within user comments can be automatically identified and analyzed using computational techniques for emotion detection. In the current study, we will represent a pipeline for emotion detection of YouTube comments for English languages. The framework based on the analysis of the YouTube comments using the MLlib library [13] of the Spark Apache framework [14], a fundamental Big Data analysis framework. Apache Spark has emerged as one of the leading frameworks in big data computing with a robust stack containing several helping tools for better Big Data handling. Based on the offered machine learning algorithms we conducted a system. In the first phase, we collected and prepared the corpus, which is composed of YouTube comments. The preparation process results of a cleaned and embedded Data. The next step is fundamental for the Dataset preparation, where we apply a clustering algorithm for the data automatic annotation. The labelling is based on the Ekman basic emotional model. The last phase is based on the labeled dataset where we selected and trained a machine learning classifiers using the MLlib library. The remainder of this paper is organized as follows. Section 2 presents the literature review and related studies. In Sections 3, we present our methodology, which includes the proposed framework. Section 4 discusses the experimental results. Finally, Section 5 presents the conclusion and perspectives.

II. THEORETICAL REFERENCE

II.1 BIG DATA

II.1.1 Big Data Definition

From a computer science perspective, data is information that can be stored and used by computer program. As a first definition, Big Data is a collection of a large and complex data that traditional data processing tools are incapable to treat. More specific technical definition, Big data is a set of data with high volume, high velocity and/or high variety that requires new forms of processing tools for improved decision making, Knowledge discovery and understanding.

II.1.2 Big Data Characteristics

To maintain a clear conceptual distinction between large data collections and big data, it is crucial to refer to a set of defining characteristics that distinguish truly massive datasets. In scientific and technological contexts, the classification of data as big data does not depend only on its size, but rather on multiple dimensions typically called Vs representing the main characteristics that enable the classification of data as Big Data or otherwise. Initially comes the volume of data, where the term Big Data essentially refers to the big size of data, emphasizing its size as a defining characteristic. In this context, volume represents the total amount or quantity of data generated by one or multiple applications. The data size directly influences the computational and storage requirements, as well as the processing strategies necessary to efficiently manage and analyze such data. The velocity takes the second position; it refers to the speed of data generation. It is a measure of how fast the data is generated and processed. The velocity plays a fundamental role in the processing challenges of Big Data.

In another perspective, Big Data is composed of a variety of data, where data is generated from multiple sources and it consists of various forms and formats. This introduces variety in data and consequently introduces 'complexity'. The variety is due to the availability of a large number of heterogeneous platforms in the industry. This means that the type to which Big Data belongs to is also an important characteristic that needs to be known for appropriate processing of data. This characteristic helps in effective use of data according to their formats, thus maintaining the importance of Big Data. Finally, the veracity is in addition considered as an important characteristic to consider the quality of data captured; it refers to the reliability and integrity of data, encompassing variations in coverage, accuracy, and relevance. Ensuring that datasets are sufficiently accurate is essential, as critical business and analytical decisions often rely on them. While many researchers believed in three and four Vs (i.e. Volume, Velocity, Variety and Veracity) for Big Data characteristics, another concept that needs to be addressed is the discussions that were taking place between researchers on the five Vs [15] and seven Vs [16] of big data.

II.1.3 Big Data Analysis

The primary objectives of Big Data analysis are resumed in two goals; the first goal is to design effective methodologies capable of accurately predicting future observations. The second goal is to extract meaningful insights into the relationships between features and response variables, thereby advancing scientific understanding. Although numerous additional Vs dimensions have been proposed in the literature, researchers consistently acknowledge Volume, Velocity, and Variety as the three fundamental characteristics of Big Data. These core attributes were initially identified in early foundational studies and continue to be emphasized as the primary dimensions

defining Big Data. Big Data are characterized by both high dimensionality and large sample sizes, which together introduce several distinctive challenges. Where the volume of data often leads to noise accumulation, spurious correlations, and incidental homogeneity, complicating accurate inference. When combined with variety of data, it further results in high computational costs and algorithmic instability, affecting model efficiency and scalability. Moreover, Big Data are characterized by the variety and velocity of data as presented in the figure 1, collected at different time points and through various technologies. These characteristics introduce experimental variations and statistical biases, thereby necessitating the development of adaptive and robust analytical methods capable of handling such complexity.

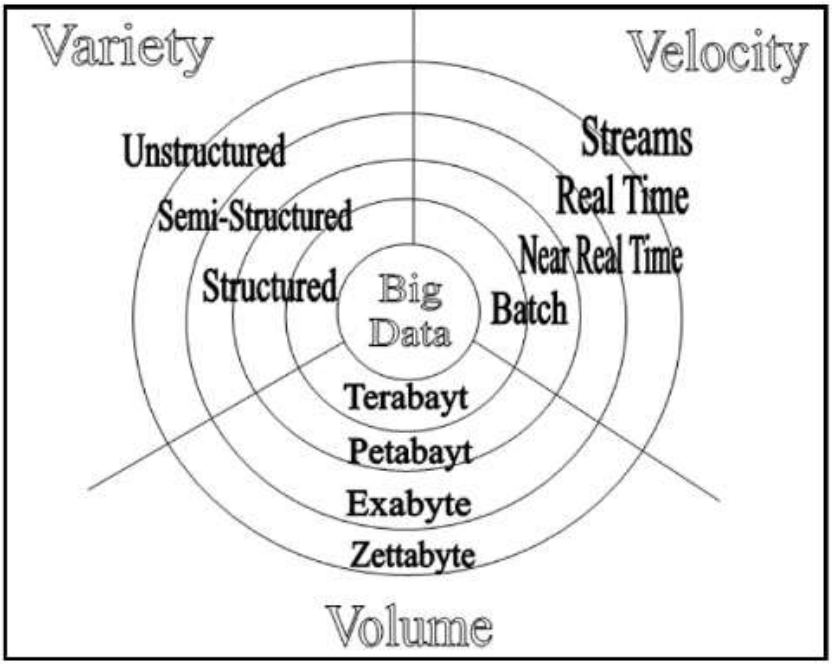


Figure 1: The three Big Data Vs.
Source: [17].

In conclusion, the different Vs characterizing Big Data need tools and techniques for data storage, applications, managements, meaning and analysis in different application domain like education and learning analytics, healthcare and biomedical research, finance and banking, social media analysis and various others interested in derive knowledge from amounts of data. In particular, data continues to grow exponentially; new technologies are evolving accordingly. Frameworks such as Hadoop, Apache Spark, and Apache Storm have gained significant popularity due to their ability to handle large-scale data processing with high efficiency. These systems are continuously advancing, leading to the development of specialized libraries such as Spark's MLlib that facilitate the integration of Machine Learning techniques within Big Data based environment. In the next section, we will introduce one of the most used frameworks designed for the process of data in the Big Data context.

II.2 APACHE SPARK

The exponential growth of data volumes in both industrial and research domains presents significant opportunities, yet also introduces significant computational challenges. As data growth have surpassed the processing capabilities of individual machines, it has become necessary to develop systems capable of distributing computations across multiple nodes. Consequently, a rise of cluster-based programming paradigms has emerged, each designed to address a variety of computational workloads. Among the most popular Big Data frameworks in recent years are: Apache Spark for distributed processing of data, Apache Kafka for real-time streaming, Hadoop HDFS is for distributed storage, Databricks is a unified Big Data and AI platform and Snowflake and Google BigQuery are top cloud analytics solutions. Initially, these paradigms were relatively specialized, with new models being introduced to meet the specific requirements of emerging application domains.

Apache Spark [18] is an open-source distributed computing framework that is designed for large scale data processing, it is considered as one of the most prominent and widely adopted cluster computing systems. Supported by a highly active open source community, Spark remains among the most dynamic projects within the Apache ecosystem figure 2. It delivers high performance for both batch and streaming workloads. This efficiency is achieved through its advanced Directed Acyclic Graph (DAG) scheduler, query optimizer, and physical execution engine. By storing data in the memory of worker nodes, Spark significantly outperforms Hadoop in scenarios that require multiple iterative operations. Apache Spark offers a more efficient and user friendly analytics environment compared to Hadoop MapReduce, enabling programs to execute up to 100 times faster in memory and up to 10 times faster on disk. Moreover, Spark provides extensive support for multiple programming languages, including Java, Python, and R, and can operate seamlessly across various environments such as Hadoop, Mesos, standalone clusters, and cloud platforms.

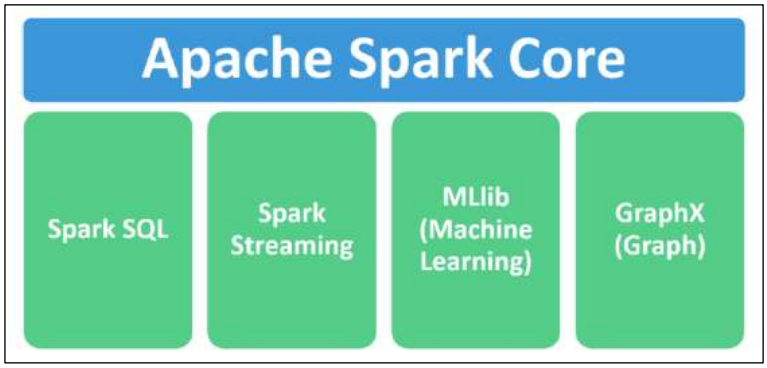


Figure 2: Apache Spark Ecosystem.
Source: Authors, (2026).

II.2.1 MLlib

MLlib [13] is a scalable, user-friendly, and comprehensive machine learning library developed as part of the Apache Spark ecosystem, offering both high performance and ease of use in distributed computing environments. It comprises a comprehensive set of machine learning algorithms, it is designed to process and analyze large-scale data in parallel. However, executing MLlib algorithms [19], [20]within Spark can present considerable technical complexity and operational effort challenges that become progressively more complex because of the intensive growth of dataset sizes.

II.2.2 Core Features of MLlib

MLlib provides a robust, flexible and comprehensive library for large-scale machine learning. Its core features include a comprehensive collection of algorithms for classification, regression, clustering, and collaborative filtering, enabling efficient analysis of massive datasets in distributed environments. Designed with scalability and usability in mind, MLlib simplifies the construction and optimization of machine learning workflows, offering a streamlined experience comparable to that of scikit-learn. Beyond algorithmic implementation, it encompasses advanced functionalities for feature extraction, selection, and scaling, as well as built in tools for model evaluation using standard performance indicators such as accuracy and precision. The library supports two main programming interfaces: the RDD-based API, which provides fine-grained control over distributed computations, and the DataFrame-based API, which enhances ease of use and performance optimization for high-level analytical tasks.

The next section introduces one of the typical fields of application of MLlib in the social big data analytics domain witch is the emotion detection in social media, more particular emotion detection from text data derived from YouTube comments.

II.3 EMOTION DETECTION FROM YOUTUBE COMMENTS

Emotions constitute an essential aspect of human life, fundamentally influencing the way individuals perceive and interpret the world. Over the past three decades, numerous approaches have been developed to advance emotion analysis ranging from manual techniques, such as psychologist-designed questionnaires, to computer-based methods. Nowadays, computer assisted emotion recognition has found widespread applications. Humans express their emotions through various modalities, including written language, vocal tone, facial expressions, physiological responses, body gestures, postures, and other physiological signals. For example, emotion detection from physiological signals [21-24] is employed in the development of smart homes and intelligent work environments. In addition, facial recognition techniques [25-27] are increasingly used in consumer services, education, and security-related domains, among others. To systematically categorize and interpret the emotional expressions, several theoretical emotion models have been proposed [28-30]. Selecting an appropriate emotion model is essential for accurately recognizing and interpreting emotions across different modalities.

Such a model should define a set of permissible emotional categories that are relevant and suitable for the specific application context or scenario. With the rapid development and evolution of social media platforms, an increasing number of users now express their emotions and opinions online in response to various events or topics, ranging from feelings of affection and agreement to dislike or sadness. Consequently, research focused on developing effective computational techniques to analyze such user-generated content and enable emotion-aware systems has garnered substantial attention from both the academic and industrial communities in recent years. In this context, analyzing the YouTube contents intended to support emotion detection objectives, has progressively evolved in recent years. Comments on YouTube videos related to a given topic often contain spontaneous emotional expressions, offering valuable insights into public sentiment.

II.4 RELATED WORKS

Emotion detection in the context of YouTube has attracted increasing scholarly attention in recent years due to the platform's massive user base and the richness of affective information embedded in user-generated content. For instance, The research presented in [31] explores the relationship between color and emotion by analyzing the rich spectrum of colors present in YouTube videos to uncover the emotional nuances they communicate. This study explores how colors influence emotional responses in movies and other visual media. By integrating classical color theory, psychological research, and modern media analysis, it seeks to uncover new ways of predicting and interpreting emotional effects in digital content. In a different perspective, recent research on YouTube video classification has shifted from purely topical categorization toward emotion based analysis, for example, the research performed in [32] have introduced models that classify videos into six basic emotions, happiness, anger, disgust, fear, sadness, and surprise, using supervised and

unsupervised learning methods, highlighting the importance of emotional information in multimedia content understanding. An other example of research in this area elaborated in [33] has examined how different modalities, text, audio, and visual features like color, interact to influence viewers emotional reactions. By analyzing textual components, including video titles, descriptions, and transcripts, alongside audiovisual cues to identify the elements that most strongly evoke emotions across various genres.

The research indicates that textual information, particularly the combination of titles, descriptions, and transcriptions, plays a central role in influencing audience emotions, highlighting the significance of linguistic context in multimodal emotion analysis. In the context of emotion analysis on YouTube comments, an extensive research has been conducted in the literature, employing a wide range of techniques and approaches to detect emotions embedded within the textual content of user comments. For example, in the study performed in [34], the authors conducted experiments on emotion classification on Indonesian YouTube comments. A corpus containing 8,115 YouTube comments is collected and manually labelled using the Ekman basic emotional model, with one added neutral label. From the experiments, they conclude that using word embedding on emotion classification task on YouTube comment can increase the accuracy of the classification and that the best method to use word embedding for the classification is by using Convolutional Neural Network algorithm. An additional work realized in [35] where the authors presented a comprehensive set of techniques to identify sentiment and extracted emotions from Bangla texts by collecting a dataset of comments from various YouTube videos.

They built deep learning based models to classify the Bangla sentences with a three class (positive, negative, neutral) and a five class (strongly positive, positive, neutral, negative, strongly negative) sentiment labels. They also built models to extract the emotion of the Bangla sentence as any one of the six basic emotions (anger, disgust, fear, joy, sadness and surprise). Finely, they evaluated the performance of their model using a new dataset of Bangla, English and Romanized Bangla comments from different types of YouTube videos. Some research has employed machine-learning technique, precisely, the supervised technique to detect emotions in YouTube comments, the work presented in [36] analyzed users expressed emotions, the study conducts a sentiment analysis based on textual comments. The authors evaluated the performance of a Long Short-Term Memory (LSTM) model on two distinct datasets, the IMDB movie reviews dataset and a YouTube video comments dataset available on GitHub, in order to examine the generalization capability and effectiveness of a single model across different but semantically related domains. The proposed LSTM model achieves good testing accuracies on the datasets, demonstrating competitive performance compared to existing sentiment analysis approaches.

The stady performed in [37] investigated the role of psychological and affective features in improving automated cyberbullying detection. It integrated personality traits, derived from the personality models, and emotional categories, based on Ekman's model extracted from textual YouTube comments. The authors evaluated whether incorporating these higher-level behavioral cues could enhance the classification effectiveness of ensemble machine learning models (Random Forest and AdaBoost), demonstrating that the fusion of personality and emotion based features gives rise to superior performance compared to traditional linguistic feature sets. Some other research has employed unsupervised approaches to identify emotional patterns in multilingual datasets, including English and Arabic comments related to social and political conflicts, highlighting the cross-linguistic and contextual complexity of emotional expression online. For example, the study performed in [38] where the main objectives were the automatic annotation of a multilingual corpus with emotional label based on the Ekman emotional model, with the investigation of the psychological impact of crises on individuals, by analyzing the emotional content embedded in YouTube comments associated with war-related videos.

The study focuses on the clustering algorithm K-means, which is an unsupervised model, used for emotion detection from YouTube comments related to the Gaza war, using data in both English and Arabic. It explores the role of Emoji's in enhancing emotional interpretation. Generally, the findings confirm the effectiveness of the clustering techniques in detecting emotional patterns from unstructured and diverse YouTube data. The proposed framework presented a scalable, language-aware solution for unsupervised emotion detection, making it particularly suitable for multilingual and sensitive contexts. Another initiative using the unsupervised techniques is presented in [39] where the authors presented there system architecture divided into two phases : at first they collected a corpus of YouTube comments from various videos, Emotion labeling is not required in this phase. In the second phase, an unsupervised machine learning approach was adopted to group or classify comments according to their emotional tone without relying on predefined emotion labels. The unsupervised machine learning algorithm was based on the work presented in [40], with modifications to improve accuracy, and utilizing YouTube comments as a rich resource for publicly available text.

In another orthogonal context, using the Big Data frameworks like Apache Spark has demonstrated significant improvement. The research performed in [41] addresses the challenge of analyzing massive volumes of audience comments on YouTube videos, where the main goal is to use machine learning algorithms in a distributed environment mainly Apache Spark to automatically analyze the large dataset of viewer comments, determine audience sentiment, and help content creators improve their strategies. The authors introduced an Improved Novel Ensemble Method (INEM), ensemble-based machine learning model that combines multiple classifiers to improve sentiment prediction. Using Spark significantly improved scalability and speed for processing and analyzing big datasets of YouTube comments. Using Twitter data the researchers in [42] address the challenge of automatically detecting emotional states expressed by users in Arabic language tweets, particularly within the context of the COVID-19 pandemic. The authors affirm that beyond simple positive/negative sentiment, fine-grained emotion classification like joy, fear, anger, sadness, disgust, surprise plays a crucial role in understanding human behavior in crises.

The aim of the research is to develop a real-time online tool that handle streaming Arabic tweets and predicts one of six basic emotions, using the Spark Apache tools to collect, pre-process, feature-extract, and apply the trained model in real time to each tweet. In the same context, the system described in [43] highlights the power of the Apache Spark platform in order to perform sentiment analysis on large volumes of data from Twitter. The motivation was that, given the volume and velocity of Twitter data streams, traditional single machine or batch approaches are insufficient. The authors built a pipeline that can scale, process the data effectively, classify tweets by sentiment, and thereby support applications such as public opinion monitoring, marketing analytics, or social behavior analysis. Using the Spark platform, machine-learning algorithms mainly: Naïve Bayes, Logistic Regression, Support Vector Machine, Decision Tree that were applied via the MLlib library, enabled efficient handling of larger tweet volumes and faster processing compared to non-distributed methods. The Table 1 summarizes the selected researches related to our study.

Table 1: Summary of prior research.

Ref.	Research Focus	Data Source	Methodology / Model	Emotion Model	Key Findings
[31]	Relationship between color and emotion in media	YouTube videos	Color spectrum analysis integrating color theory, psychology, and media analysis	Plutchik emotional model (8 emotions)	Colors strongly influence emotional response; visual features contribute to emotional interpretation.
[32]	Emotion-based video classification	YouTube videos	Supervised & unsupervised learning	Ekman emotional model (Six basic emotions)	Models effectively classify videos into six emotional categories, emphasizing the role of emotional context in multimedia understanding.
[33]	Multimodal emotion analysis (text, audio, visual)	YouTube videos	Feature fusion: textual (titles, descriptions, transcripts) + audiovisual	Russell and Mehrabian model	Textual information, especially combined metadata, plays a dominant role in predicting audience emotions from video.
[34]	Emotion classification on Indonesian YouTube comments	YouTube comments	Word embeddings + CNN	Ekman emotional model + Neutral	Word embeddings improve accuracy; CNN performs best for emotion classification.
[35]	Emotion and sentiment detection in Bangla YouTube comments	Bangla, English, Romanized Bangla YouTube comments	Deep-learning models	Sentiment (3-class, 5-class)&Ekman emotions	Created multilingual datasets; models perform well for both sentiment and emotion tasks.
[36]	Cross-domain emotion analysis (YouTube & IMDB)	YouTube comments & IMDB reviews	LSTM (supervised)	Sentiment analysis	LSTM generalizes well across domains; competitive performance on multiple datasets.
[37]	Cyberbullying detection enhanced by emotion & personality traits	YouTube comments	Ensemble ML (Random Forest, AdaBoost)	Ekman + Personality traits	Fusion of personality and emotion features improves detection accuracy beyond linguistic baselines.
[38]	Multilingual unsupervised emotion detection (English & Arabic)	War-related YouTube comments	K-Means clustering (unsupervised)	Ekman emotional Model	Effective at uncovering emotional patterns; emojis enhance interpretation; suited to multilingual contexts.
[39]	Unsupervised emotion detection for automatic datasets annotation	YouTube comments	Unsupervised algorithm based on [40]	Ekman emotional Model	Automatically clusters comments by emotional tone; works without predefined labeled dataset.
[41]	Distributed emotion analysis using Big Data frameworks	Large-scale YouTube comments	Apache Spark + Improved Novel Ensemble Method (INEM)	Sentiment analysis	Spark ensures scalability & speed; ensemble model boosts predictive accuracy on massive datasets.
[42]	Real-time Arabic emotion detection during COVID-19	Twitter data	Apache Spark pipeline	Ekman emotional Model	Real-time detection of six emotions; Spark supports live streaming and online analysis.
[43]	Scalable sentiment analysis on social-media data streams	Twitter data	Spark MLlib (Naïve Bayes, LR, SVM, DT)	Sentiment analysis	Distributed Spark framework enables fast, large-scale sentiment classification.

Source: Authors, (2026).

III. MATERIALS AND METHODS

III.1 DATASET CREATION

The dataset is created by the extraction of comments from two videos using YouTube API version 3.0. We manually select these videos from the CNN and BBC news channels in English language based on their popularity (number of views and number of comments) dated respectively 2025 and 2021. We limit the number of comments for each video up to **4490**, which results a total number of **8980** comments written in English. Table 2 provides a description of our dataset features.

Table 2: Dataset description.

Dataset features	Number of comments
Total number of comments	8980
Comments with Emojis	932
Comments without Emojis	8048

Source: Authors, (2026).

The dataset annotation is based on the Ekman emotional model where six basic emotion, which defines six fundamental emotion categories: happiness, sadness, fear, disgust, anger, and surprise. The annotation process is mainly based on the pipeline described by the work defined in [38] where we used the K-means model which is an unsupervised machine learning technique. Based on the same strategy, we performed a preprocessing pipeline to prepare the data for the labeling process.

III.2 DATASET PREPARATION

The Dataset preparation process begins after the data collection, which results a corpus of unlabeled YouTube comments described in the section above. The resulting Dataset is annotated by the K-means clustering model based on the Ekman's basic emotional model. The figure 3 illustrates the different steps of the dataset annotation process. Data preprocessing enhances text quality and consistency through cleaning and normalization. It ensures cleaner and more consistent textual input for emotion detection. It removes noise (URLs, mentions, symbols), normalizes text (repetitions, hashtags, whitespace, case), and eliminates stop words to improve model accuracy. Numerical values are retained, as they may hold meaningful contextual or emotional significance. After cleaning, text tokenization is performed using BERT tokenizer, which splits text into sub word units and converts them into token IDs for model input. This enables BERT's bidirectional encoding to capture full contextual meaning which is crucial for emotion detection. The high dimensional output from BERT increased computational complexity and noise during clustering. To avoid this issue, Principal Component Analysis (PCA) was applied to reduce dimensionality while preserving key variance and data structure, resulting in a more efficient and effective clustering process.

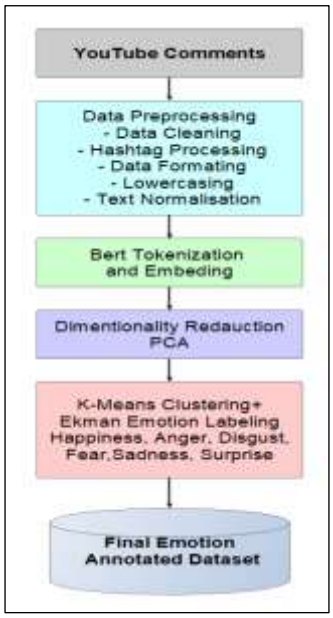


Figure 3: Data annotation Pipeline.
Source: Authors, (2026).

III.3 CLUSTERING TECHNIQUE

Unsupervised learning techniques provide an effective way for textual data annotation with minimal human intervention, motivating the use of clustering for automatic text annotation. The annotation of the YouTube comments Dataset is assured by using a clustering technique basically K-means, which efficiently handles large datasets and balanced clusters, to assign emotional labels to YouTube comments. This approach produces a labeled dataset that forms the basis for training supervised models for accurate emotion detection. In the next section we will present the selected collection of machine learning techniques where the main objective is using the resulting labeled dataset after the clustering step in the training for preparing the model to the emotion prediction and recognition.

III.4 MACHINE LEARNING TECHNIQUES

In this work, we utilized three classification algorithms in order to implement the emotion detection system. At this level, the selected classifiers are chosen from the machine learning models defined in the MLlib [13] Spark's library. We selected a collection composed of classic models where the objective is to compare their accuracy and decide the best model in the context of emotion

recognition from amount volume of text specifically in the context of YouTube comments analysis. The selected model are basically the Naïve Bayes, Logistic Regression and Random Forest classifier.

III.4.1 Naive Bayes

The Naïve Bayes classifier is a probabilistic multiclass classification algorithm grounded in Bayes' theorem. In this model, each data instance is represented as a feature vector, under the assumption that all features are conditionally independent given the class label. The algorithm is computationally efficient, requiring only a single pass through the training data to estimate parameters. During training, the conditional probability distribution of each feature with respect to each class is computed. Subsequently, Bayes' theorem is applied to determine the most probable class label for unseen instances.

III.4.2 Logistic Regression

The Logistic Regression is a statistical classification model in which the dependent variable can assume one of several discrete values. It employs the logistic function to model the relationship between the input features and the probability of belonging to a specific class. While the model is primarily designed for binary classification, it can be extended to handle multiclass problems using different strategies.

III.4.3 The Random Forest Classifier

The Random Forest algorithm is an ensemble learning technique based on the bagging principle. It addresses the main limitation of a single decision tree by constructing multiple decision trees on randomly selected subsets of the dataset. The final prediction is then obtained by taking the majority vote of the individual trees. The Random Forest Classifier in Apache Spark's MLlib is an ensemble-learning model based on the Random Forest algorithm, where his primary role is tasks classification. It is based on the construction of a multitude of decision trees during training where it returns as result the class with the highest vote among the individual trees. In the next section, we will present the different obtained results with a discussion of the most relevant among them.

III.5 EVALUATION METRICS

The performance of the used machine learning algorithms is measured using a set of evaluation metrics, in the first range the accuracy is used; it is the most widely used. For many years, accuracy served as the basic criterion for comparing machine learning models effectiveness. However, it is not sufficient alone to make an accurate decision of the model performance. Accordingly, we used the Precision, the Recall and the F-score for more precise performance measurement of the used models. Accuracy measures the overall correctness of predictions, while Precision and Recall evaluate the model's ability to correctly identify relevant classes. The F1-score provides a harmonic mean between Precision and Recall, providing a more robust performance indicator in imbalanced datasets. It ranges from zero to one where a higher F1-score indicates better model performance.

IV. RESULTS AND DISCUSSIONS

In this section, we present and analyze the experimental results obtained, along with an evaluation of the effectiveness of the tree used algorithms (Logistic Regression, Naïve Bayes, and Random Forest) from the MLlib Spark's library applied to the resulting labeled dataset. In addition, we discusses the performance of the proposed models based on various evaluation metrics. The goal is to evaluate the effectiveness and overview capability of each algorithm on the selected dataset by analyzing and comparing the obtained results. To evaluate model performance, we used standard metrics notably: Accuracy, Precision, Recall, and F1-score. Table 3 presents the performance of the three used classification models based on the four standard evaluation metrics presented above. As illustrated, the Logistic Regression and Naïve Bayes models achieved the highest accuracy values, demonstrating superior performance to Random Forest Classifier by a considerable margin.

Table 3: Performance results for the three algorithms.

Models	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9289805269186713	0.9293267627553385	0.9289805269186712	0.9282860057563744
Naïve Bayes	0.8178694158075601	0.8370928837933002	0.8178694158075601	0.8217422717081797
Random Forest Classifier	0.5664375715922108	0.49976124982922965	0.5664375715922108	0.41087645030132036

Source: Authors, (2026).

For instance, the Logistic Regression obtained an accuracy of 92.3%, with a precision of 91.8% and an F1-score of 92.0%, indicating consistent performance across classes. Overall, Logistic Regression achieved the best performance across all metrics, with an accuracy of approximately 0.93 and an F1-score of 0.93, indicating consistent and reliable classification results. Naïve Bayes followed, obtaining an accuracy of 0.82 and an F1-score of 0.82, which suggests satisfactory performance but slightly lower precision and recall compared to Logistic Regression. In contrast, the Random Forest Classifier recorded the lowest performance, with an accuracy of 0.57 and an F1-score of 0.41. This poor performance may be due to overfitting, suboptimal parameter tuning, or an insufficient number of estimators, which limited its generalization ability on the dataset. In summary, the comparative results indicate that Logistic Regression provides the most effective balance between accuracy, precision, and recall for this emotion detection task, outperforming both Naïve Bayes and Random Forest. The Figure 4 illustrates the visualization via a bar chart comparing the performance metrics of the three models. We can clearly notice that Logistic Regression outperforms the other models across all metrics (Accuracy, Precision, Recall, and F1-Score), while the Random Forest Classifier shows the weakest performance overall.

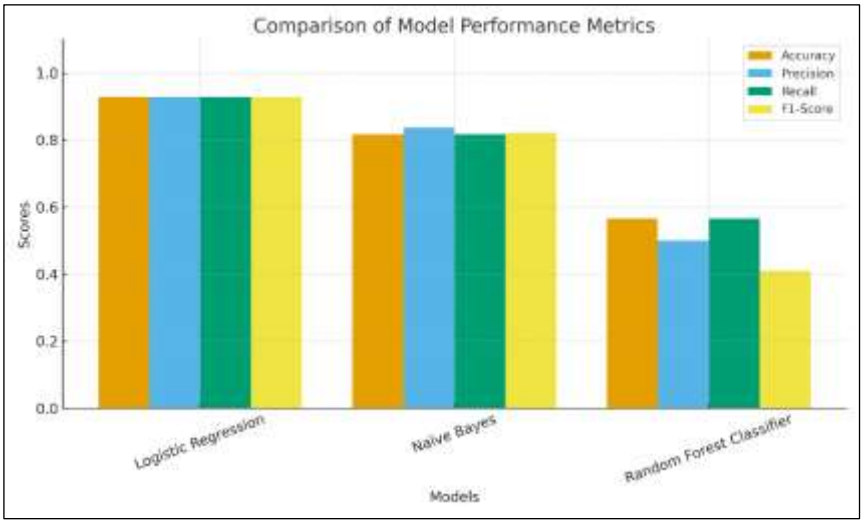


Figure 4: Performance comparison of the three models.
Source: Authors, (2026).

V. CONCLUSIONS

In this study, we presented our framework designed for the emotion detection from YouTube comments, leveraging the machine learning capabilities of the Spark's Apache MLlib library. We proposed a scalable emotion detection framework for analyzing YouTube comments using Apache Spark large-scale abilities. The integration of Machine Learning algorithms with Natural Language Processing techniques enabled efficient processing and classification of emotions across large dataset. By employing K-Means clustering for automatic corpus annotation and supervised classifiers trained on the labeled dataset, the system demonstrated improved accuracy and scalability compared to traditional approaches. The experimental evaluation demonstrated that the framework could efficiently process large volumes of data while maintaining high performance, where the Logistic Regression surpassed 92% in both metrics, accuracy and F1-Score, making it a promising solution for real-world social media analytics. Moreover, the study underlines the importance of scalable architectures in Big Data driven emotion analysis. Future research will focus on the integration of deep learning architectures and multimodal data (text, audio, and video) to improve the precision and expressiveness of emotion detection systems in complex online environments. In the same context, the work can be adopted for steaming data using the Spark Steaming library for Emotion detection tasks in large steaming social data.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Wafa Saadi, Fatima Zohra Laallam and Messaoud Mezati.
Methodology: Wafa Saadi.
Investigation: Wafa Saadi.
Discussion of results: Wafa Saadi, Fatima Zohra Laallam and Messaoud Mezati.
Writing – Original Draft: Wafa Saadi**.**
Writing – Review and Editing: Wafa Saadi.
Resources: Wafa Saadi.
Supervision: Fatima Zohra Laallam and Messaoud Mezati.
Approval of the final text: Wafa Saadi, Fatima Zohra Laallam and Messaoud Mezati.

VIII. REFERENCES

[1] A. Badshah, A. Daud, R. Alharbey, A. Banjar, A. Bukhari, and B. Alshemaimri, 'Big data applications: overview, challenges and future', Artif. Intell. Rev., vol. 57, no. 11, p. 290, Sep. 2024, doi: 10.1007/s10462-024-10938-5.

[2] R. Kamal and P. Saxena, Big Data Analytics. Mcgraw Hill Education, 2019.

[3] M. Ianni, E. Masciari, and G. Sperlí, 'A survey of Big Data dimensions vs Social Networks analysis', J. Intell. Inf. Syst., vol. 57, no. 1, pp. 73–100, Aug. 2021, doi: 10.1007/s10844-020-00629-2.

[4] W. G. Parrott, Emotions in Social Psychology: Essential Readings. in Key readings in social psychology. Psychology Press, 2001. [Online]. Available: https://books.google.dz/books?id=FNBnKLvFdM4C

[5] R. W. Picard, Affective Computing. The MIT Press, 1997. doi: 10.7551/mitpress/1140.001.0001.

[6] P. Ekman, 'Basic Emotions', in Handbook of Cognition and Emotion, 1st ed., T. Dalgleish and M. J. Power, Eds., Wiley, 1999, pp. 45–60. doi: 10.1002/0470013494.ch3.

[7] R. Plutchik, 'A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION', in Theories of Emotion, Elsevier, 1980, pp. 3–33. doi: 10.1016/B978-0-12-558701-3.50007-7.

[8] R. Plutchik and H. Kellerman, Theories of emotion. in Emotion, theory, research, and experience, no. v. 1. New York: Academic Press, 1980.

[9] H. Zhang, ‘A Study of Human Emotion Analysis Based on Social Media’, in Proceedings of the 2023 2nd International Conference on Social Sciences and Humanities and Arts (SSHA 2023), vol. 752, M. F. B. Sedon, I. A. Khan, M. C. BiRkök, and K. Chan, Eds., in Advances in Social Science, Education and Humanities Research, vol. 752. , Paris: Atlantis Press SARL, 2023, pp. 174–180. doi: 10.2991/978-2-38476-062-6_23.

[10] K. A. S, M. S, and R. R. L, ‘Emotional Analysis on Social Media Platform’, in 2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), Chennai, India: IEEE, Mar. 2025, pp. 1–6. doi: 10.1109/ICDSAAI65575.2025.11011895.

[11] J. M. Garcia-Garcia, V. M. R. Penichet, and M. D. Lozano, ‘Emotion detection: a technology review’, in Proceedings of the XVIII International Conference on Human Computer Interaction, Cancun Mexico: ACM, Sep. 2017, pp. 1–8. doi: 10.1145/3123818.3123852.

[12] A. Saxena, A. Khanna, and D. Gupta, ‘Emotion Recognition and Detection Methods: A Comprehensive Survey’, J. Artif. Intell. Syst., vol. 2, no. 1, pp. 53–79, 2020, doi: 10.33969/AIS.2020.21005.

[13] ‘MLlib: RDD-based API - Spark 4.0.1 Documentation’. Accessed: Oct. 20, 2025. [Online]. Available: https://spark.apache.org/docs/latest/mllib-guide.html

[14] ‘Apache SparkTM - Unified Engine for large-scale data analytics’. Accessed: Oct. 20, 2025. [Online]. Available: https://spark.apache.org/

[15] T. L. Nguyen, ‘A Framework for Five Big V’s of Big Data and Organizational Culture in Firms’, 2018 IEEE Int. Conf. Big Data Big Data, pp. 5411–5413, 2018.

[16] M. A. Khan, M. F. Uddin, and N. Gupta, ‘Seven V’s of Big Data understanding Big Data to extract value’, in Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education, 2014, pp. 1–5. doi: 10.1109/ASEEZone1.2014.6820689.

[17] S. Sagiroglu and D. Sinanc, ‘Big data: A review’, in 2013 International Conference on Collaboration Technologies and Systems (CTS), 2013, pp. 42–47. doi: 10.1109/CTS.2013.6567202.

[18] M. Zaharia et al., ‘Apache Spark: a unified engine for big data processing’, Commun. ACM, vol. 59, no. 11, pp. 56–65, Oct. 2016, doi: 10.1145/2934664.

[19] X. Meng et al., ‘MLlib: Machine Learning in Apache Spark’, J. Mach. Learn. Res., vol. 17, no. 34, pp. 1–7, 2016.

[20] L. Theodorakopoulos, A. Karras, and G. A. Krimpas, ‘Optimizing Apache Spark MLlib: Predictive Performance of Large-Scale Models for Big Data Analytics’, Algorithms, vol. 18, no. 2, p. 74, Feb. 2025, doi: 10.3390/a18020074.

[21] X. Li et al., ‘EEG Based Emotion Recognition: A Tutorial and Review’, ACM Comput. Surv., vol. 55, no. 4, pp. 1–57, Apr. 2023, doi: 10.1145/3524499.

[22] L. Wang, J. Hao, and T. H. Zhou, ‘ECG Multi-Emotion Recognition Based on Heart Rate Variability Signal Features Mining’, Sensors, vol. 23, no. 20, p. 8636, Oct. 2023, doi: 10.3390/s23208636.

[23] J. Perdiz, G. Pires, and U. J. Nunes, ‘Emotional state detection based on EMG and EOG biosignals: A short survey’, in 2017 IEEE 5th Portuguese Meeting on Bioengineering (ENBENG), Coimbra, Portugal: IEEE, 2017, pp. 1–4. doi: 10.1109/ENBENG.2017.7889451.

[24] Aloy Anuja Mary G, ‘Emotion Detection Through Electrocardiogram Signal Classification in an IOT Environment with Deep Neural Networks’, J. Electr. Syst., vol. 20, no. 3, pp. 1620–1630, May 2024, doi: 10.52783/jes.3657.

[25] Prof. K. Sarvakar, ‘A Survey Of Face Emotion Recognition Using Deep Learning Methods’, Jun. 03, 2025, In Review. doi: 10.21203/rs.3.rs-6794812/v1.

[26] H. A. Shehu, W. N. Browne, and H. Eisenbarth, ‘Emotion categorization from facial expressions: A review of datasets, methods, and research directions’, Neurocomputing, vol. 624, p. 129367, Apr. 2025, doi: 10.1016/j.neucom.2025.129367.

[27] H. Liu, ‘Emotion Detection through Body Gesture and Face’, 2024, arXiv. doi: 10.48550/ARXIV.2407.09913.

[28] K. J. Lakshmi Bai, A. Harshita, T. Gaddam, S. Mirtipati, and D. Peyyala, ‘A Comprehensive survey on Emotion Recognition’, in 2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG), Indore, India: IEEE, Dec. 2024, pp. 1–6. doi: 10.1109/ICTBIG64922.2024.10911870.

[29] A. R. Murthy and K. M. Anil Kumar, ‘A Review of Different Approaches for Detecting Emotion from Text’, IOP Conf. Ser. Mater. Sci. Eng., vol. 1110, no. 1, p. 012009, Mar. 2021, doi: 10.1088/1757-899X/1110/1/012009.

[30] R. A. García-Hernández et al., ‘A Systematic Literature Review of Modalities, Trends, and Limitations in Emotion Recognition, Affective Computing, and Sentiment Analysis’, Appl. Sci., vol. 14, no. 16, p. 7165, Aug. 2024, doi: 10.3390/app14167165.

[31] M. C. Cakmak, M. Shaik, and N. Agarwal, ‘Emotion Assessment of YouTube Videos using Color Theory’, in Proceedings of the 2024 9th International Conference on Multimedia and Image Processing, Osaka Japan: ACM, Apr. 2024, pp. 6–14. doi: 10.1145/3665026.3665028.

[32] Y.-L. Chen, C.-L. Chang, and C.-S. Yeh, ‘Emotion classification of YouTube videos’, Decis. Support Syst., vol. 101, pp. 40–50, Sep. 2017, doi: 10.1016/j.dss.2017.05.014.

[33] N. Yousefi, M. C. Cakmak, and N. Agarwal, ‘Examining Multimodel Emotion Assessment and Resonance with Audience on YouTube’, in Proceedings of the 2024 9th International Conference on Multimedia and Image Processing, Osaka Japan: ACM, Apr. 2024, pp. 85–93. doi: 10.1145/3665026.3665039.

[34] J. Savigny and A. Purwarianti, ‘Emotion classification on youtube comments using word embedding’, in 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA), Denpasar: IEEE, Aug. 2017, pp. 1–5. doi: 10.1109/ICAICTA.2017.8090986.

[35] N. Irtiza Tripto and M. Eunus Ali, ‘Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments’, in 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), Sylhet: IEEE, Sep. 2018, pp. 1–6. doi: 10.1109/ICBSLP.2018.8554875.

[36] P. Sharma, S. Gupta, P. Gangwar, P. Kumar, and N. Saxena, ‘Decoding emotions: LSTM-based sentiment analysis for movie reviews and YouTube comments’, presented at the 4TH INTERNATIONAL CONFERENCE ON INNOVATION IN IOT, ROBOTICS AND AUTOMATION (IIRA 4.0), Moradabad, India, 2025, p. 020044. doi: 10.1063/5.0246918.

[37] V. Balakrishnan and S. K. Ng, 'Personality and emotion based cyberbullying detection on YouTube using ensemble classifiers', Behav. Inf. Technol., vol. 42, no. 13, pp. 2296–2307, Oct. 2023, doi: 10.1080/0144929X.2022.2116599.

[38] W. Saadi, F. Z. Laallam, M. Mezati, C. R. HALIMI, and E. H. BRIKI, 'Emotion Detection in Conflict Discourse: An Unsupervised Approach to Multilingual YouTube Comments', J. Inf. Syst. Eng. Manag., vol. 10, no. 58, pp. 7–15, 2025.

[39] D. Yasmina, M. Hajar, and A. M. Hassan, 'Using YouTube Comments for Text-based Emotion Recognition', Procedia Comput. Sci., vol. 83, pp. 292–299, 2016, doi: 10.1016/j.procs.2016.04.128.

[40] A. Agrawal and A. An, 'Unsupervised Emotion Detection from Text Using Semantic and Syntactic Relations', in 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Macau, China: IEEE, Dec. 2012, pp. 346–353. doi: 10.1109/WI-IAT.2012.170.

[41] K. Subha and N. Bharathi, 'Spark-based Big Data Sentiment Analysis of Social Media Comments', in 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India: IEEE, Nov. 2023, pp. 164–169. doi: 10.1109/ICSCNA58489.2023.10370518.

[42] N. Abdelhady, I. E. Elsemman, and T. H. A. Soliman, 'A real-time predicting online tool for detection of people's emotions from Arabic tweets based on big data platforms', J. Big Data, vol. 11, no. 1, p. 171, Nov. 2024, doi: 10.1186/s40537-024-01035-z.

[43] A. Baltas, A. Kanavos, and A. K. Tsakalidis, 'An Apache Spark Implementation for Sentiment Analysis on Twitter Data', in Algorithmic Aspects of Cloud Computing, T. Sellis and K. Oikonomou, Eds., Cham: Springer International Publishing, 2017, pp. 15–25.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 222-235. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2874

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

HYBRID GPT AND NEURAL MODELS FOR PERSONALIZED E-COMMERCE: A NOVEL FRAMEWORK FOR ADAPTIVE AND TRANSPARENT PRODUCT RECOMMENDATIONS

Pavan Gunda[1], Thirupathi Rao Komati[2]

[1, 2]Department of Computer Science and Engineering, GITAM School of Technology, GITAM, Deemed-to-be University, Visakhapatnam-530045, AP, India

[1]https://orcid.org/0009-0008-6163-6789, [2]https://orcid.org/0000-0002-0907-6824

Email:*gsraghupavan9@gmail.com, tkomati@gitam.edu

ARTICLE INFO

Article History
Received: October 30, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
E-commerce,
Explainability,
Product Recommendation,
GPT Models,
Hybrid Similarity.

ABSTRACT

In order to transform e-commerce customization, the study presents a product recommendation system that makes use of OpenAI's GPT model's better resources in conjunction with neural integration. Traditional systems often rely on collaborative filtering or content-based models. This may lack precision and adaptability to individual needs. The proposed method overcomes these limitations by integrating hierarchical classification. Hybrid similarity scoring and GPT-based annotation capabilities provide more precise and interpretable rules. The multi-step technique ensures that the machine leads to different types of statistics which includes textual product descriptions and generated metadata to gain meaningful insights. This is because pre-processing techniques determine the specific format of numbers and facts,while embeddings built from neural models provide rich semantic knowledge about relationships between products. The weighted scoring mechanism dynamically adjusts to the choices made by individuals. This increases the machine's capability to generalize to the needs of a man or woman. In addition, GPT-inspired factors promote consideration and transparency in explaining the underlying reasons in natural language. The results of the experiment validated the device's ability to offer advanced customization and user satisfaction. By combining domain-specific data with existing IA strategies, this structure sets a new benchmark for precision-driven, people-centric ecommerce applications.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

ProductGPT is a revolutionary concept that harnesses the power of artificial intelligence to transform the product comparison and recommendation landscape. At the heart of ProductGPT lies the integration of ChatGPT, a cutting-edge language model that enables the generation of human-like explanations and bridges the gap between mathematical metrics and natural language. This synergy empowers ProductGPT to provide users with personalized, transparent, and explainable product recommendations, thereby redefining the user experience. ChatGPT plays a pivotal role in ProductGPT by generating intuitive explanations that elucidate the reasoning behind product recommendations. By decoding complex mathematical metrics, such as cosine similarity and Euclidean distance, ChatGPT translates technical jargon into accessible language, enabling users to make informed decisions. Furthermore, ChatGPT enhances personalization by incorporating user preferences and feedback, ensuring that product recommendations are tailored to individual needs. The fusion of ChatGPT and mathematical metrics in ProductGPT gives rise to a robust and user-centric system. By bridging the gap between mathematics and natural language, ProductGPT provides a unique and innovative approach to product comparison and recommendation. As we delve into the specifics of ChatGPT's role in ProductGPT, we will explore how this integration revolutionizes the product recommendation landscape, enabling users to navigate complex product landscapes with ease and confidence.

I.1 CHATGPT'S ROLE IN GENERATING EXPLANATIONS

ChatGPT plays a key role in making the recommendation system not only functional, but also useful and it can also be interpreted [1]. It provides personalized explanations for product recommendations. The system processes complex mathematical calculations, such as similarity scores obtained by combining products. Classification, price, and category, but these are abstract and not intuitive for users. With ChatGPT integration, these calculation results are translated into user-friendly narratives [2]. For example, the get_explanation function asks ChatGPT for the following basic inputs: Product name. Product resources (such as prices, discounts, ratings) and user profile settings. This information is used to manage personalized descriptions, such as "This product was recommended because of a deep discount. Corresponds to the price range you want and is in the desired category." Such narratives ensure users understand how their preferences $U_p = \{C, R, P, D\}$ where C is category, R is rating, P is price, and D is discount — are matched with product features$P_f = \{C', R', P', D'\}$. By bridging computational reasoning with natural language, ChatGPT enhances transparency, trust, and user satisfaction [3].

I.2 DECODING MATHEMATICAL METRICS WITH CHATGPT

The recommendation system relies on mathematical metrics such as cosine similarity. Euclidean distance and mixed integration to calculate the similarity between user preferences and product resources [4]. For example, hybrid coalescence similarity is calculated as follows:

$$S_{hybrid} = \frac{1}{1+||E_1 - E_2||} . S_p \tag{1}$$

Where E_1 and E_2 are product embeddings, $||E_1 - E_2||$ are their Euclidean distance, and S_p is the price similarity score, because this mathematical method guarantees accuracy. As a result, it lacks interpretability for non-technical users. ChatGPT acts as an interpreter, translating these scores into understandable language. For example, ChatGPT can trigger a description such as "This product is highly recommended, because they are very similar to the selected category. Offer attractive prices and have great reviews from other customers." This change allows users to make informed decisions while appreciating the complex calculations behind them.

I.3 ENHANCING PERSONALIZATION AND USER EXPERIENCE

ChatGPT customizes the recommendation process by describing or aligning two product resources with user-defined needs [5]. Assume a user profile U_p specify preferences for products with minimum rating of R_u, maximum price of P_u, and target discount of D_u. The recommendation system evaluates the weighted similarity score:

$$S_{total} = w_c \cdot S_c + w_p \cdot S_p + w_r \cdot S_r + w_e \cdot S_e \tag{2}$$

Where $w_c, w_p, w_r, and\ w_e$ are weights of the category, price, rating and embedding. ChatGPT sets the context for this score by presenting a story such as: "This product meets user needs with ratings up to 4.5, huge discounts of 30%, and prices within a certain range." By aligning numerical results with human language, ChatGPT helps ensure users feel that recommendations are tailored [6]. To meet their specific needs, this greatly increases engagement and satisfaction. Additionally, the ability to explain instructions in detail increases confidence in the system's decisions.

I.4 BRIDGING MATHEMATICS AND NATURAL LANGUAGE WITH CHATGPT

Integrating ChatGPT into the recommendation system bridges the gap between mathematical rigor and user accessibility. [7] For example, similarity between product embeds is calculated using cosine similarity:

$$Cosine\ Similarity = \frac{\vec{A}.\vec{B}}{||\vec{A}||.||\vec{B}||} \tag{3}$$

Where, $\vec{A}, \vec{B}$ represents product feature vectors. Additionally, numerical features such as price and classificationare normalized and compared using techniques such as Euclidean distance[8]. These calculations create quantitative insights. Because its technical nature can alienate users, ChatGPT translates these insights into relevant stories, such as "This product aligns closely with your needs due to its excellent reviews, competitive price and relevance to the chosen category." This combination of computational precision and linguistic clarity is an example of a strong, user-centred IA system [9]. Users are guaranteed to benefit from the complex algorithms without being overwhelmed by complexity.

II. LITERATURE SURVEY

[10] has proposed a hybrid NLP methodology for fine-tuning of LLM to identify the accurate culture products. Especial the data was chosen in symbolic semantic info which displays as cultural symbols; these are created in three phases: physical, ethereal, and behavioural. For clear understanding culture images for GPT 3.5 approach an additional data was gathered from Palace Museum. Uncoded symbols are designed for couple of idea makings design thoughts and other was decoded with full design.

To clarify these design principles a high level common language is utilized. For these designs a cultural meaning was designed with high clarifiers. By processing the model all symbols are included to the training and while extraction the outcome was changed to

meaningful cultures. This data-driven approach seeks to link historical cultural knowledge with modern product design. [11] couple of approaches are designed for extracting creation phases and optimization of product ads with LLM and Multi-model info. The data includes write up and visual info such as images and metrics, which creates search engines as user friendly. Initial stage of processing data to model and pre-process the data with suitable image was reflected from the collected data. Then, LLMs are improved so that they can understand this data better and give detailed statements that take into account the user needs and industry trends. This model is mainly utilized with GPT-3.5 and 4. This approach also utilizes NLP for the extraction of meaningful words for detailing, SEO was improved. As the description was generated the appropriate match was generated an image with related to product and it makes user friendly.

[12] suggested method is based on automatically summarising studies that look at how food affects people from NDA review papers. It uses LLM including GPT-4 to make short outlines that are based on facts through a repeated asking approach. There are three steps to this process. First, the model makes a first report of a food effect report at an advanced level. In the second step, keyword-focused prompts make sure that important pharmacokinetic measures like AUC, Cmax, and Tmax are considered. Lastly, a question that limits the length of the answer makes sure that the summary is short, usually no more than two or three lines. The human reader and LLM work together over and over to make the results better and better, both in terms of truth and readability. To compare the success of GPT-4 and ChatGPT, evaluations use both automatic measures ROUGE and evaluations by FDA professionals. To reduce manual PSG development for generic medicine evaluation, FDA personnel will examine credible, draft-quality summaries.

[13] have introduced an established method to get trait values from e-commerce product descriptions that aren't organised using LLMs. The goal is to raise the AVE of the output by using the zero-shot and few-shot features of LLMs, especially when there isn't a lot of marked training data available. They show different prompt forms for teaching LLMs how to do AVE jobs. These include written and JSON-based versions of the goal model. We are using the GPT-3.5, GPT-4, and Llama-3 models in the tests to see how well these reminder templates work in both zero-shot and few-shot situations. The way looks at how well LLMs can pull trait values without a lot of task-specific training and compares them to cutting-edge PLMs like BERT. The F1-score is one of several evaluation measures used to check how well the models can correctly extract attribute values, even for attributes that haven't been seen yet. Additionally, the way looks into what happens when you add example values, do in-context learning demos, and fine-tune the models.

[14] integrates LLM with graph-based data to improve proposal systems. The system aims to make it easier to predict how a person and an item will connect by mixing the text-generating abilities of LLMs with the relationship modelling abilities of GNNs that use a graph structure. The main strategy is creating an attentive LLM backbone with a graph structure that records both first- and second-order relationships between people and objects. This way uses GNNs to get edge data within nodes, like people and things, and add this connection information to the LLM's attention system. To help the model understand the many linkages inside the recommendation system, pre-training is carried out utilising crowd-contextual stimuli, such as user and item descriptions and interactions. The model is then fine-tuned using personalised predictive hints to make it even better at making accurate, personalised suggestions. By combining graph data and written data, the method gives a fuller picture of user tastes and item characteristics, which lets it make more accurate suggestions.

[15] a novel use of an LLM ensemble technique for e-commerce product attribute value retrieval. The writers suggest an ensemble framework to take advantage of the fact that different LLMs, like GPT, PaLM, and Llama, have different strengths and weaknesses when it comes to different jobs. Using crowd sourcing methods as inspiration, the ensemble approach gives weights to the results of several LLMs based on how well they guess the values of product attributes. The method improves and speeds up the process of assigning weights by using the Dawid-Skene Approach, which is often used with crowd-labeled data. The ensemble method takes the results from all the models and gives more weight to the LLMs that do a better job while reducing the weight of the models that do a worse job. By utilizing unorganized text the product has achieve high accuracy and retrieved reliable data related to products. Walmart's internal datasets showed enhanced "age" and "gender" extraction using the approach.

[16] one new way to use LLM to improve e-commerce results. Before, search engines used organised data, which meant that they could index and retrieve data from sources that were not structured, like customer reviews. To change this, the study suggests turning structured and semi-structured data, like product catalogues and categories, into written explanations that can be used with LLM training. To train or fine-tune LLMs, the system uses texts with encoded product IDs that have been marked up. This lets them act as a single input for search and selection jobs. With this combination, the model can use domain-specific knowledge and give solutions that are relevant to the situation. According to the design, the framework will combine databases with LLM features, which will allow for the instant recovery of changing data like prices and stock levels. Creating unique IDs for database items, using templates to make labelled texts, and using self-supervised learning to finetune the model are all parts of the approach.

[17] improves product recommendation systems using K-means clustering, CBF, and hierarchical clustering and GPT-4. These algorithms learn patterns and provide personalised suggestions on product datasets, which GPT-4 evaluates. Beyond assessment, GPT-4 uses natural language comprehension to improve product feature semantics and recommendation accuracy. To simplify implementation, e-commerce operators may use CSV files for training and distributing suggestions using a flask-based API. This API emphasises non-technical usability with model training & real-time predictions. GPT-4 automates review using binary ratings and reasoning, boosting cost and scalability over human evaluation. E-commerce companies may customise models for their datasets while preserving efficiency and relevance using the adaptable technique. Our framework improves unsupervised learning and LLM integration for scalable, accurate, & easy to use product system recommendations.

Table 1: Analysis on Existing Systems.

Author	Algorithm	Merits	Demerits
Yang Yin et al	GPT-3.5, PLM	Fine-tuning the model for automatic retrievals	Utilized dataset contains less content
Satish Kathiriya et al	LLM	Accuracy and discovery of items are enhances for user experiences	Computation performance was not up to the mark for real-time apps
Yiwen Shi et al	Iterative prompting with GPT4 & ChatGPT	It improves the understanding of drugs name by iterating the text	Token size was limited
Alexander Brinkmann et al	LLM	Holds more training data with efficiency and robustness	Comparison performance does not have huge differences
Xinyuan Wang et al	Integrate LLM	By combining two orders the prediction was efficient	Need more memory for storing paths
Chenhao Fang et al	LLM-Ensemble	Values for attributes are extracted efficiently were it works accurately even offline	Little time-consuming while extracting dynamical learning weights
Haixun Wang et al	LLM	Conversion of structure to unstructured data was efficient	Cost-Efficient
Konstantinos I. Roumeliotis et al	CBF, Cutting-edge gpt4LLM, K-means Cluster, Hierarchical Clustering	Compared to all approaches CBF has achieved high accuracy	Overfitting

Source: Authors, (2026).

III. PROPOSED METHODOLOGY

The product recommendation system constructed the usage of GPT showcases several modern functions that distinguish it from traditional structures. The integration of OpenAI's GPT for producing motives brings a brand new level of personalization and transparency to suggestions, enabling users to recognize why precise merchandise has been suggested. Unlike traditional collaborative filtering or content-based models, this machine introduces a hybrid similarity metric that blends class-based totally matching, charge-discount alignment, and neural embeddings for textual product descriptions. The use of embeddings permits the device to understand nuanced relationships between merchandise, even when they lack common functions. Furthermore, the gadget consists of hierarchical category extraction (e.g., high-degree, secondary-level categories) for refined filtering. Another novelty lies inside the weighted similarity scoring mechanism, which may be dynamically adjusted to in shape user choices or domain-precise desires. By incorporating diverse similarity metrics, the model ensures extra precise and diverse recommendations. The interpretability of the improved version through GPT bridges the distance between machine-generated decisions and human interpretability. This makes it a modern machine for contemporary e-commerce systems.

III.1 NOVELTY

The proposed model represents a significant improvement over the old approaches. Although older versions may rely on traditional similarity measures such as cosine similarity to raw numerical features or category matching. But the new model suggests a hybrid approach that combines neural input for textual data. This lets in for a deeper understanding of product descriptions and relationships. Additionally, the hierarchical separation of lessons inside the new model allows for multi-level filtering. This becomes no longer present in the antique device. GPT integration is every other top notch thing for interpretation. This makes guidelines more obvious and intuitive. Additionally, a brand new weighted similarity metric offers flexibility to spotlight sure functions primarily based on user preferences, Whilst the old model Static or much less precise calculations may be used. The machine also includes extra similarity signs inclusive of price discounts and scores. The similarities make the new device more accurate, versatile and compliant with present day e-commerce requirements, together with solving the restrictions of the old gadget greater effectively.

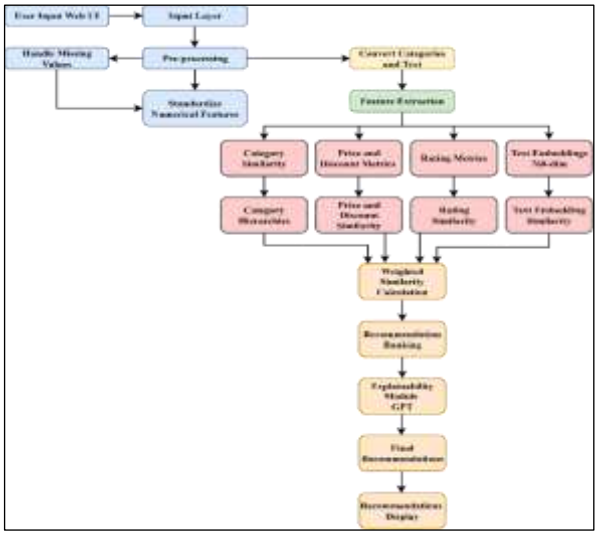


Figure 1: Block Diagram for Proposed Recommendation System.
Source: Authors, (2026).

III.2 DATA PRE-PROCESSING

Pre-processing helps ensure that datasets are clean, consistent, and ready for advanced analysis [18]. Missing values in numeric columns are handled by defining the median value:

$$x_{filled} \equiv \begin{cases} xifx \neq NaN \\ median(x)ifx = NaN \end{cases} \quad (4)$$

Categorical columns are imputed with the mode:

$$c_{filled} \equiv \begin{cases} cifc \neq NaN \\ mode(c)ifc = NaN \end{cases} \quad (5)$$

To standardize numerical data for cosine similarity, we apply:

Where z is the value after standardization, x is the actual value, μ is the mean, and σ is the standard deviation. For textual fields, embeddings (E) are generated using neural models:

$$E = f_{embedding}(T) \quad (6)$$

Where $f_{embedding}$ is a neural network-based function (e.g., BERT or GPT embeddings), and T is the product description [19].These embeddings allow the model to efficiently encode the semantic relationships between products. When combined with hierarchical class separation, these steps ensure data consistency and provide a robust model for the input format. It helps make predictions more accurate and to improve user experience.

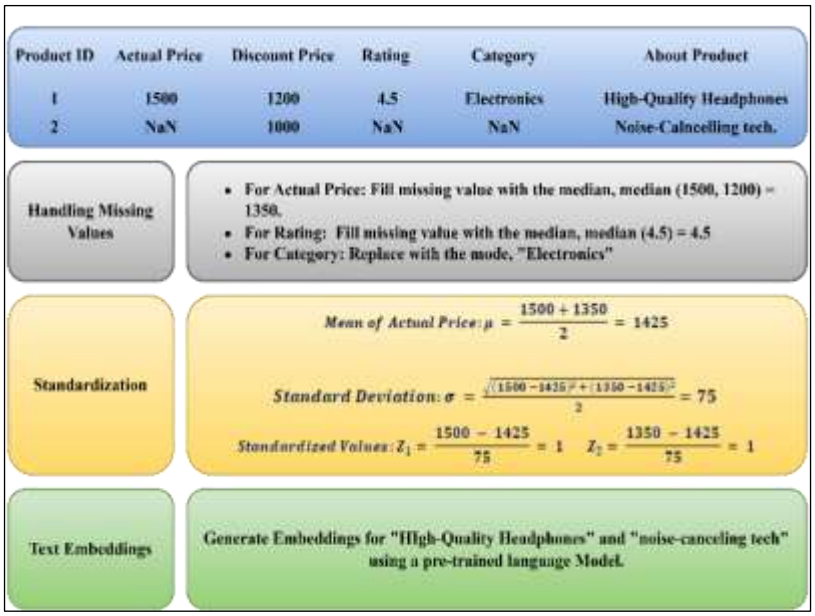


Figure 2: Pre-processing of Elements.
Source: Authors, (2026).

III.3 FEATURE EXTRACTION AND HIERARCHICAL CATEGORY MANAGEMENT

In the approach, the next step after preparing the data is to extract the categories, focusing on hierarchical categories [20]. The system divides the category space into high-level categories, secondary level and tertiary-level categories in function separator *get_high_level_category(category)* using the delimeter'|'. The first section is selected as the high level category. Mathematically, if the category string is as specified c so higher level classes are:

$$high_(level_categroy(c)\ ) = split(c," \mid ")[0] \quad (7)$$

For the secondary and tertiary levels, similar logic applies:

$$secondary_level_category(c) = split(c," \mid ")[1](if\ exists) \quad (8a)$$

$$tertiary_level_category(c) = split(c," \mid ")[2](if\ exists) \quad (8b)$$

These extracted items are then added to the dataset as new categories. This allows the model to consider product distribution at different levels. This structure is useful for more granular filtering and recommendations [21]. Because items in higher groups are grouped and sorted more efficiently, a hierarchical structure allows the model to identify patterns in the data based on different levels of variance. Further improve the quality of recommendations by including deeper relationships between constructs.

Table 2: Hypothetical Input Data.

Product ID	Category
101	Electronics | Smartphones | Accessories
102	Home | Furniture | Living Room
103	Electronics | Laptops

Source: Authors, (2026).

Extraction Logic

$$high_level_categroy(c) = split(c,|)[0]$$

$$secondary_level_category(c) = split(c," \mid ")[1](if\ exists)$$

$$tertiary_level_category(c) = split(c," \mid ")[2](if\ exists)$$

Extracted Categories

1. For Product 101:
- High-level: Electronics
- Secondary-level: Smartphones
- Tertiary-level: Accessories

2. For Product 102:
- High-level: Home
- Secondary-level: Furniture
- Tertiary-level: Living Room

3. For Product 103:
- High-level: Electronics
- Secondary-level: Laptops
- Tertiary-level: None (no 3rd segment)

Table 3Final Table.

Product ID	High Level	Secondary Level	Tertiary Level
101	Electronics	Smartphones	Accessories
102	Home	Furniture	Living Room
103	Electronics	Laptops	

Source: Authors, (2026).

Let's go through more detailed numerical examples for each of the five steps with expanded calculations which ensure to show all intermediate steps and formulas.

III.4 FEATURE EXTRACTION AND HIERARCHICAL CATEGORY MANAGEMENT

Table 4: Hypothetical Input Data.

Product ID	Category
101	Electronics | Smartphones | Accessories
102	Home | Furniture | Living Room
103	Electronics | Laptops

Source: Authors, (2026).

Extraction Logic:

$$high_level_categroy(c) = split(c,|)[0]$$

$$secondary_level_category(c) = split(c," \mid ")[1](if\ exists)$$

$$tertiary_level_category(c) = split(c," \mid ")[2](if\ exists)$$

Extracted Categories

1. For Product 101:
- High-level: Electronics
- Secondary-level: Smartphones
- Tertiary-level: Accessories

2. For Product 102:
- High-level: Home
- Secondary-level: Furniture

- Tertiary-level: Living Room

3. For Product 103:
- High-level: Electronics
- Secondary-level: Laptops
- Tertiary-level: None (no 3rd segment)

Table 5: Final Table.

Product ID	High Level	Secondary Level	Tertiary Level
101	Electronics	Smartphones	Accessories
102	Home	Furniture	Living Room
103	Electronics	Laptops	

Source: Authors, (2026).

III.5 STANDARDIZATION OF NUMERICAL FEATURES

Standardization of numerical features is an important step to make the variables comparable by converting them to a common scale. The numerical features in the dataset, such as *actual_price*, *discounted_price*, *discount_percentage*, *rating*, and *rating_count*, are standardized using the formula for Z-score normalization:

$$z = \frac{x-\mu}{\sigma} \tag{9}$$

Where x is the original value, μ is the mean, and σ is the standard deviation of the feature. The purpose of this adjustment is to ensure that all numeric are standardized, so that the similarity measure is not affected by large differences between features and able to work effectively. This standard is important for similarity - the base models participate equally in the calculations. After setting the standard the dataset is more suitable for similarity-based analysis. This leads to more accurate recommendations. Normalization reduces bias caused by disproportionately large or small values in a data set. This will improve the generalizability of the model.

III.6 TEXT EMBEDDING GENERATION

Another important feature in this example is the text entry for the field *about_product*. Embeddings are mathematical representations of meaningful data results. The function *get_product_embedding (product_description)* tracks the process of creating an embedding. If T stands for product specification, Embedding E is obtained from a pre-trained model, e.g. BERT or GPT embedding generation formula can be specified as follows:

$$E(T) = f_{embedding}(T)$$

Where $f_{embedding}$ is a neural network-based function (e.g., BERT or GPT embeddings), and T is the product description. These embeddings permit the machine to compare the semantic similarity of different merchandise; the closer pals are in vector space. The extra semantically similar the products can be, the greater similar they will be. This approach allows the gadget to advise products that don't monitor genuine characteristics however are comparable inside the text description. This makes the model greater flexible and might locate hidden relationships among products. In addition to information, it is able to additionally make use of rich semantic context.

Table 6: Hypothetical Input Data.

Product ID	About Product
101	"High-quality smartphone accessories."
102	"Comfortable and modern living room furniture."
103	"Latest laptops with powerful specifications."

Source: Authors, (2026).

Using a pre-trained embedding model (OpenAI), generate embeddings.

Table 7: Embedding.

Product ID	Embedding (E)
101	[0.45, 0.60, 0.32, 0.11]
102	[0.12, 0.35, 0.55, 0.40]
103	[0.80, 0.70, 0.40, 0.25]

Source: Authors, (2026).

III.7 HYBRID SIMILARITY COMPUTATION

The subsequent step includes calculating the similarity between products based on multiple capabilities: class, fee, cut price, score, and embeddings. The similarity among two products is computed the usage of an aggregate of numerous metrics. For class similarity, the formula is straightforward: if each merchandise belongs to the same high-degree category, the similarity is 1, in any other case 0:

$$category_similarity(p1,p2) = \begin{cases} 1 \ if \ high_{level_{category(p1)}} = high_level_category(p2) \\ 0 \ otherwise \end{cases} \quad (10)$$

For price and discount similarity, the difference in prices and discount percentages is normalized to create a similarity score:

$$price_{discount_{similarity(p_1,p_1)}} = \frac{1}{1+|price(p_1)-price(p_2)|+|\alpha-\beta|} \quad (11)$$

Where α is $discount_percent(p_1)$ and β is $discount_percent(p_2)$. For rating similarity, a similar approach is used:

$$rating_similarity(p_1,p_2) = \frac{1}{1+|rating(p_1])-rating(p_2)|} \quad (12)$$

For embedding similarity, the Euclidean distance between the embeddings of two products is used:

$$embedding_similarity(p_1,p_2) = \frac{1}{1+||E(p_1)-E(p_2)||} \quad (13)$$

Finally, these individual similarities are combined using weighted summation to generate a final similarity score:

$$total_{similarity(p_1,p_2)} = w_1.category_{similarity(p_1,p_2)} + w_2.price_{discount_{similarity(p_1,p_2)}} + w_3.rating_similarity(p_1,p_2) + w_4.embedding_similarity(p_1,p_2) \quad (14)$$

Where, w_1, w_2, w_3, w_4 are the weights assigned to each feature's similarity. The model then uses this final similarity score to rank the products and recommend those most aligned with the user's preferences.

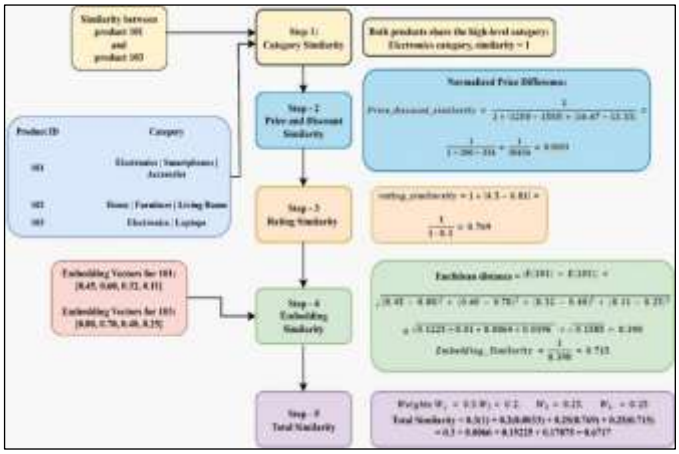


Figure 3: Similarity Computations.
Source: Authors, (2026).

III.8 FINAL RECOMMENDATION AND EXPLAINABILITY

After calculating the similarity ratings, the goods are ranked in descending order of similarity. The top tips are selected for display to the user. The machine also carries GPT-based totally explainability. For every encouraged product, the machine generates a textual explanation as to why the product turned into recommended, the use of the formula:

$$explanation = g_{GPT}(product_{features}, user_{profile}) \quad (15)$$

Where g_{GPT} is OpenAI GPT model that generates explanation primarily based at the product capabilities and the consumer profile. This enhances the transparency and trustworthiness of the recommendation device. Finally, the gadget returns the pinnacle N hints, whole with their factors, pictures, and hyperlinks for further exploration. The hybrid similarity model combined with explainability affords a comprehensive advice framework that caters to the consumer's choices and enhances the buying revel in. By presenting explanations for every advice, the model fosters extra person believe and engagement, for that reason enhancing the general consumer relevant.

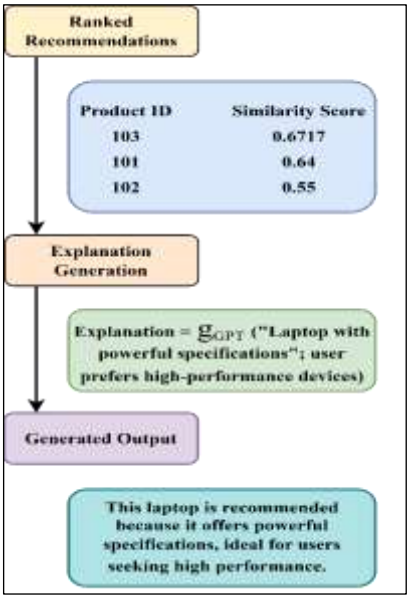


Figure 4: Explainable AI Recommendation.
Source: Authors, (2026).

IV. RESULTS AND DISCUSSIONS

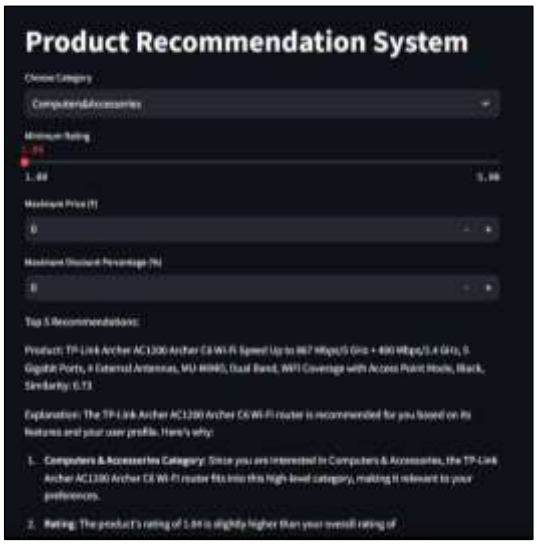


Figure 5: User Interface of the Application.
Source: Authors, (2026).

The Figure 5 shows the main user interface of the product recommendation application created using Streamlit. The interface is designed to be intuitive and easy to use. It has a header that clearly states the purpose of the application. Users will see several interactive elements such as drop-down menus, sliders, and number entry fields. These components allow users to specify their preferences, such as selecting categories, minimum score required, maximum price and maximum discount percentage. The clean layout ensures that users can easily navigate the interface without any confusion. Interactive widgets provide real-time feedback. It makes the user experience smooth and attractive. The modular interface design reflects careful consideration in structuring user input. This screenshot shows the application's main function of storing user preferences in an orderly manner. It emphasizes the simplicity and accessibility of the design. To support both novice and experienced users, clear division of elements ensures that each input field is easily identified to prevent ambiguity during interactions.

The Figure 6 offers an outline of the product classes available in the dataset. The classes are displayed in a dropdown menu, allowing customers to select the excessive-degree class that suits their preferences. These categories are derived from the ‘class’ column in the dataset and include numerous product kinds. The dropdown capability enables dynamic filtering of hints based totally at the person’s preference. Users can fast navigate thru the listing and make selections that replicate their shopping pursuits. The inclusion of categories allows slender down the search space, making hints greater focused and relevant. This feature not handiest enhances user delight but additionally improves the efficiency of the advice engine. The design guarantees that even a huge variety of categories are presented in a prepared and reachable way. By permitting category-specific filtering, the utility guarantees that users acquire consequences aligned with their number one hobbies. This screenshot highlights how the software leverages established statistics to improve personalization and person experience.

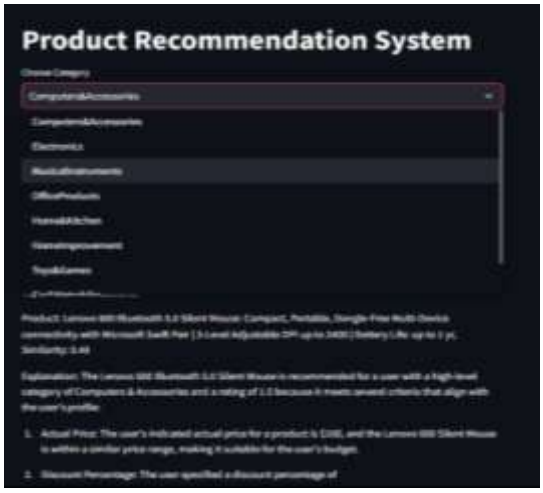


Figure 6: Categories Available.
Source: Authors, (2026).

In Figure 7, the input fields are highlighted, showing how users can customize their search for recommendations. These inputs include sliders for specifying the minimum acceptable product rating and number fields for defining the maximum price and discount percentage. By combining these inputs with the category selection, the application gathers detailed user preferences to create a comprehensive user profile. This profile serves as the foundation for calculating similarity scores and filtering products. The flexibility of the input options ensures that users can fine-tune their preferences to match specific needs.

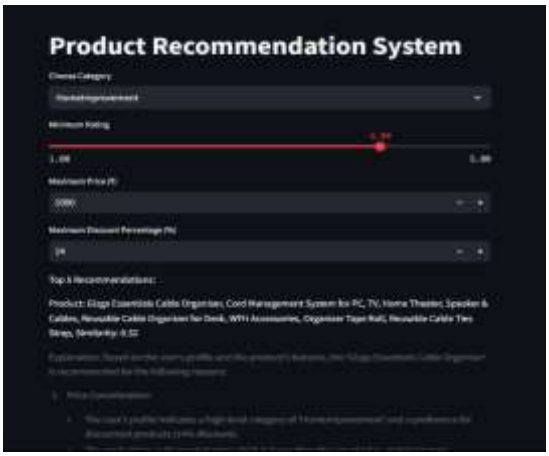


Figure 7: Input features given to the Application.
Source: Authors, (2026).

For instance, the sliders provide granular control over rating thresholds, while the number inputs allow precise budget constraints. This level of customization makes the application versatile and accommodating to a wide range of use cases. The intuitive placement of input fields ensures that users can quickly configure their preferences without requiring additional guidance. Overall, this screenshot illustrates the application's commitment to delivering a tailored user experience by empowering users to define their own criteria for recommendations.



Figure 8: Top 5 recommendations generated.
Source: Authors, (2026).

The Figure 8 showcases the consequences of the advice set of rules, showing the top 5 recommended products. Each product is indexed with its call and similarity score, indicating how well it fits the user's preferences. The guidelines are looked after in descending order of similarity, making sure that the most applicable alternatives seem on the top. This view underscores the effectiveness of the similarity metrics used inside the utility, along with category similarity, fee-bargain alignment, and score alignment. The concise presentation of effects makes it smooth for customers to assess their alternatives at a glance. Additionally, the inclusion of similarity rankings provides transparency into the recommendation procedure, constructing accept as true with the customers. The visible readability of the guidelines web page ensures that customers can fast become aware of their desired picks and proceed to explore further info. This screenshot reflects the utility's capacity to synthesize consumer inputs into actionable outputs, highlighting the performance and precision of the underlying set of rules.



Figure 9: Product Details, Similarity Percentage, Explanation, Product Image and Product Link for the recommended product.
Source: Authors, (2026).

The Figure 9 provides a detailed view of a recommended product. It consists of critical records which include the product's name, similarity percentage, and a brief clarification of why it turned into encouraged. Additionally, the product's picture and a clickable link to the product page are displayed. The explanation is generated the use of OpenAI's GPT model, adding a layer of transparency to the recommendation system through describing how the product suits the consumer's choices. The inclusion of visuals and direct links enhances consumer engagement, bearing in mind seamless exploration and buying selections. The similarity percentage gives insight into how intently the product aligns with the person's inputs, while the reason gives context for the recommendation. This mixture of functions ensures that users have all the important information to make knowledgeable selections. The utility's potential to provide complete statistics in a visually attractive manner enhances its usability and effectiveness. This screenshot encapsulates the utility's attention on handing over fee through targeted and customized recommendations.



Figure 10: Forwarded to the product page when clicked on provided Link.
Source: Authors, (2026).

The Figure 10 demonstrates the functionality of the clickable product links, displaying how customers are redirected to the respective product pages. Once a person clicks on a product link, they're taken to the e-commerce platform wherein they are able to view extra details, study critiques, and complete their purchase. This function bridges the space between pointers and motion, permitting users to act on their options with minimal attempt. The easy redirection procedure highlights the practical software of the utility, streamlining the online buying enjoy whilst making sure person satisfaction. The integration of clickable links guarantees that customers can transition seamlessly from exploring hints to making purchasing selections. This screenshot also emphasizes the utility's function as a facilitator in the e-commerce adventure, simplifying the technique of coming across and obtaining favoured merchandise. By supplying a direct course to the product page, the utility eliminates needless steps, enhancing the overall efficiency and comfort of the consumer enjoy.

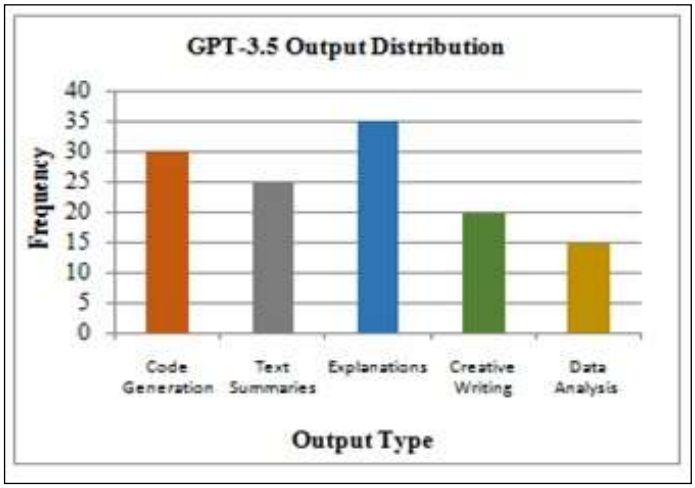


Figure 11: GPT-3.5 Output Distribution.
Source: Authors, (2026).

The above Figure 11 presents a breakdown of the different types of content generated by GPT-3.5, highlighting its versatility in various applications. Among the five categories: Code Generation, Text Summaries, Explanations, Creative Writing, and Data Analysis, the highest frequency is observed in Explanations, suggesting that GPT-3.5 is extensively used to clarify complex topics, making it a valuable tool for education, research, and professional learning. Code Generation follows closely, indicating the model's effectiveness in assisting programmers by producing functional and efficient code snippets. Text Summaries also constitute a significant portion of GPT-3.5's output, showing its ability to condense large volumes of information into concise, readable formats, particularly useful for summarizing reports, academic papers, or news articles. While GPT-3.5 can generate imaginative content and perform data-driven tasks, Creative Writing and Data Analysis are relatively less frequent, indicating that these areas are not its primary strengths or most common use cases. Overall, the distribution reflects GPT-3.5's capability to adapt to multiple domains, with a strong emphasis on technical and educational applications.

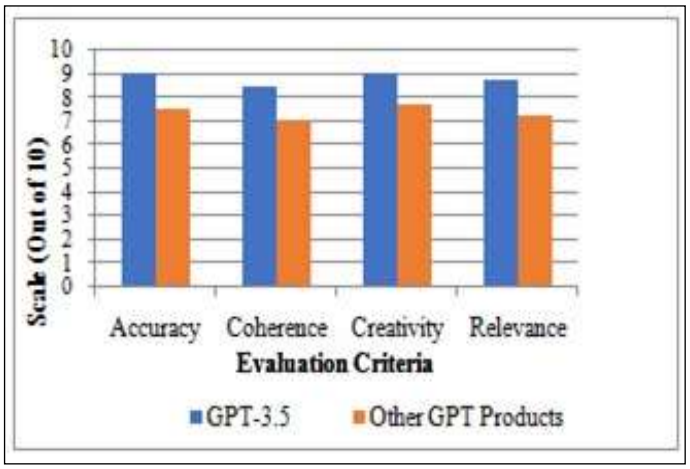


Figure 12: Comparision of GPT-3.5 Performance.
Source: Authors, (2026).

The above Figure 12 compares GPT-3.5's performance against other GPT models, based on four evaluation criteria: Accuracy, Coherence, Creativity, and Relevance. GPT-3.5 consistently outperforms other models across all criteria, demonstrating its superiority in generating high-quality responses. It scores significantly higher in Accuracy, reflecting its ability to provide factually correct and reliable information. GPT-3.5 also excels in Coherence, ensuring logically structured and contextually appropriate responses, and scores well in Creativity, producing original and engaging content. Furthermore, it shows a strong advantage in Relevance, providing responses aligned with user intent and offering meaningful insights. Overall, the graph highlights GPT-3.5's superior linguistic and analytical capabilities, making it a preferred choice for applications requiring precision, clarity, and contextual understanding, and demonstrating how improved AI models can significantly enhance the quality of generated content across multiple domains.

V. CONCLUSIONS

The proposed recommendation system effectively integrates modern AI techniques. To meet the challenges of traditional e-commerce platforms by leveraging the interpretive power of OpenAI's GPT model and the structural insights of hierarchical class separations, the machine thus achieves an excellent balance between precision and scalability. The hybrid similarity scoring mechanism enhances the version's ability to distinguish different relationships between products while neural integration provides a richness of meaning that is lacking in conventional similarity measures. The dynamic weighting of similarity measures guarantees flexibility and scalability.

Meet the needs of many people and various space requirements. Incorporating GPT-based annotation capabilities not only improves the user experience but also improves user experience, but also add consideration Provide clear and interpretable reasons for each recommendation. This transparency promotes consumer engagement and loyalty. Filling a critical gap in traditional recommendation structures, the device's modular and scalable design also allows for perfect integration into modern ecommerce workflows. By combining data based on semantic text analysis, the structure can therefore provide a precise, versatile, and actionable dictionary. Future trends may focus on increasing scalability. Reduce computational costs and integrate real-time observation mechanisms. This groundbreaking technique creates a solid foundation for the next generation of user-centred intelligent recommendation systems.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Pavan Gunda, Thirupathi Rao Komati.
Methodology: Pavan Gunda, Thirupathi Rao Komati.
Investigation: Pavan Gunda, Thirupathi Rao Komati.
Discussion of results: Pavan Gunda, Thirupathi Rao Komati.
Writing – Original Draft: Pavan Gunda, Thirupathi Rao Komati.
Writing – Review and Editing: Pavan Gunda, Thirupathi Rao Komati.
Resources: Pavan Gunda, Thirupathi Rao Komati.
Supervision: Pavan Gunda, Thirupathi Rao Komati.
Approval of the final text: Pavan Gunda, Thirupathi Rao Komati.

VII. REFERENCES

[1] Kumar, A., &Lakkaraju, H. "Manipulating large language models to increase product visibility", 2024, doi:10.48550/ARXIV.2404.07981

[2] Shahin, M., Chen, F.F., Hosseinzadeh, A. et al. "A novel approach to voice of customer extraction using GPT-3.5 Turbo: linking advanced NLP and Lean Six Sigma 4.0", Int J AdvManufTechnol 131, 3615–3630 (2024). https://doi.org/10.1007/s00170-024-13167-w

[3] Jagatap, A., Merugu, S, & Comar, P. M, "Improving search for new product categories via synthetic query generation strategies", 2024, May 13, Companion Proceedings of the ACM Web Conference 2024, 29–37. Presented at the WWW '24: The ACM Web Conference 2024, Singapore, doi:10.1145/3589335.3648299

[4] Cheng, Y., Chen, J., Huang, Q., Xing, Z., Xu, X., & Lu, Q. "Prompt Sapper: A LLM-empowered production tool for building AI chains", 2024, ACM Transactions on Software Engineering and Methodology, 33(5), 1–24. doi:10.1145/3638247

[5] Das, S., Saboo, R., Vadrevu, C. S. K., Wang, B., & Xu, S. "Applications of LLMs in E-commerce search and product knowledge graph: The DoorDash case study", 2024, March 4, Proceedings of the 17th ACM International Conference on Web Search and Data Mining. Presented at the WSDM '24: The 17th ACM International Conference on Web Search and Data Mining, Merida Mexico. doi:10.1145/3616855.3635738

[6] Feng, P., He, Y., Huang, G., Lin, Y., Zhang, H., Zhang, Y., & Li, H. "AGILE: A novel reinforcement learning framework of LLM agents", 2024, doi:10.48550/ARXIV.2405.14751

[7] P. Gunda and T. R. Komati, "Enhancing User Experience and Mitigating BIAS in Electronic Device Recommendations on e-Commerce Platforms: Leveraging NLP and GPT Technologies," 2024 4th International Conference on Sustainable Expert Systems (ICSES), Kaski, Nepal, 2024, pp. 149-156, doi: 10.1109/ICSES63445.2024.10763266

[8] Roumeliotis, K. I, Tselikas, N. D, & Nasiopoulos, D. K, "LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation", 2024, Natural Language Processing Journal, 6(100056), 100056. doi:10.1016/j.nlp.2024.100056

[9] Erlansyah, D. Mukminin, A, Julian, D, Negara, E. S, Aditya, F, & Syaputra. R, "Large language model (LLM) comparison between GPT-3 and PaLM-2 to produce Indonesian cultural content", 2024, Eastern-European Journal of Enterprise Technologies, 4(2 (130)), 19–29. doi:10.15587/1729-4061.2024.309972

[10] Yin, Y, Ding, S, Zhang, X, Wang, C, Li, X, & Cai, R. "Cultural Product Design Concept Generation with Symbolic Semantic Information Expression Using GPT", 2024, June 24, https://doi.org/10.21606/drs.2024.508

[11] Kathiriya, S, Mullapudi, M, & Karangara, R, "Optimizing ECommerce Listing: LLM Based Description and Keyword Generation from Multimodal Data", 2023, Article in International Journal of Science and Research, https://doi.org/10.21275/ SR24304113521

[12] Shi, Y, Ren, P, Wang, J, Han, B, Valizadeh Aslani, T, Agbavor, F, & Liang, H, "Leveraging GPT-4 for food effect summarization to enhance product-specific guidance development via iterative prompting", 2023, Journal of biomedical informatics, 148, 104533.

[13] Brinkmann, A, Shraga, R, & Bizer, C, "ExtractGPT: Exploring the Potential of Large Language Models for Product Attribute Value Extraction",2023 http://arxiv.org/abs/2310.12537

[14] Wang, X, Wu, L, Hong, L, Liu, H, & Fu, Y, "LLM-Enhanced User-Item Interactions: Leveraging Edge Information for Optimized Recommendations", 2024, http://arxiv.org/abs/2402.09617

[15] Fang, C, Li, X, Fan, Z, Xu, J, Nag, K., Korpeoglu, E., Kumar, S., & Achan, K. "LLM-Ensemble: Optimal Large Language Model Ensemble Method for E-commerce Product Attribute Value Extraction", 2024, SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2910–2914. https://doi.org/10.1145/3626772.3661357

[16] Wang, H., & Na, T. "Rethinking E-Commerce Search", In ACM SIGIR Forum, 2024, January, Vol. 57, No. 2, pp. 1-19. New York, NY, USA: ACM.

[17] Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. "Precision-Driven Product Recommendation Software: Unsupervised Models, Evaluated by GPT-4 LLM for Enhanced Recommender Systems", 2024, Software, 3(1), 62–80. https://doi.org/10.3390/software3010004

[18] GUNDA, Pavan & Komati, Thirupathi Rao, “Integrating self-attention mechanisms for contextually relevant information in product management”, 2024, International Journal of Computational and Experimental Science and Engineering, 10(4). doi:10.22399/ijcesen.651

[19] Li, P, Castelo, N, Katona, Z, & Sarvary, M, “Frontiers: Determining the validity of Large Language Models for automated perceptual analysis”, 2024, Marketing Science, 43(2), 254–266. doi:10.1287/mksc.2023.0454

[20] Ilagan, J. B, Ilagan, J. R, Zulueta, P. Y, Rodrigo, M. M, “Optimizing Conversational Commerce Involving Multilingual Consumers Through Large Language Models’ Natural Language Understanding Abilities”, 2024, In: Degen, H., Ntoa, S. (eds) Artificial Intelligence in HCI. HCII 2024. Lecture Notes in Computer Science(), vol 14736. Springer, Cham.https://doi.org/10.1007/978-3-031-60615-1_4

[21] S. G and M. A. K A, "Empowering Product Purchase Decisions through AI-Driven Review Analysis and Summarization," 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2024, pp. 155-162, doi: 10.1109/ICICT60155.2024.10544601

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 236-243. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2876

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

INVESTIGATION OF THE EFFECT OF NANOPARTICLES ON ENGINE PERFORMANCE AND EMISSION CHARACTERISTICS IN DIESEL ENGINES

Sathiyaraj.S*[1], Imthiyas.A[2], Kumaran.P[3], Silamburajan.M[4], Rengaraju.G[5]

[1, 2, 3,4,5]Department of Mechanical Engineering, Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research Foundation (DU), Tamil Nadu, India.

[1]https://orcid.org/0000-0001-8504-0412, [2]https://orcid.org/0009-0002-1516-9426, [3]https://orcid.org/ 0000-0003-2781-8999
[4]https://orcid.org/0009-0008-1628-1472, [5]https://orcid.org/0009-0002-2055-1400

Email: *sathiyaraj@avit.ac.in, imthiyas@avit.ac.in, kumaranp@avit.ac.in,silambu5995@gmail.com, rengarajummp97@gmail.com

ARTICLE INFO

Article History
Received: October 31, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Gongura Seed Biodiesel,
TiO2,
BTE,
NOx,
Carbon Monoxide,
Hydrocarbon

ABSTRACT

To overcome the global power crisis, alternatives to fossil fuels, whose reserves are dwindling at an alarming rate, must be discovered. We can make use of a wide variety of plants, including trees, shrubs, and herbs, to produce biodiesel. The oxidation stability and cold filter plugging point of dual biodiesel are the cold flow properties that are discussed in this paper. The best sample is taken and put through its tempo in the diesel engine having variable compression ratio to determine biodiesel blends' performance and emission characteristics. The experimental examination of the engine's performance and emission characteristics. The parameters of the engine evaluated and (i) Carbon Monoxide (CO), (ii) Hydrocarbon (HC), (iii) Nitrogen Oxide (NOX) were analyzed, also Brake Thermal Efficiency (BTE), Specific fuel consumption (SFC) and emission were compared.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The world is presently accosted with two major challenges: The Fuel energy crises and pollution of ecosystem. The huge variations in crude prices are determined internationally according to the market demands, and current consumption rates limit petroleum reserves to nearly 40 years. For the last 50 years, numerous countries throughout the world have been evolving new crops to boost the biomass substitute base for bio-energy production. Presently India is importing nearly 65% petroleum from other countries and the fluctuating price of crude scattered the Indian economy. Next to the U.S.A and China, India is anticipated to become the next highest consumer of fuel in 2025, with consumption increasing at a 7.5% annual rate. 85% of oil imported is used for transportation and energy production, and the country's economy is heavily dependent on fossil fuel imports. Due to the above, the position of bio fuel as a transportation fuel will be significant. The use of bio oils as a replacement for diesel fuel was established in 1900 by Rudolph Diesel who invented Diesel engine when vegetable oil was considered as an alternative. The use of Oil as diesel fuel was limited as it has high viscosity which is about 10 times the fossil fuels.

The properties of vegetable oils have to be transformed in order to become accustomed of the fuel to existing engines. Numerous vegetable oil derived products has projected as an additional fuel for diesel engines ASTM has established specific standards for biodiesel under ASTM D6751, which specifies the requirements for biodiesel made from vegetable oils, animal fats, and other renewable feedstocks. This standard ensures that biodiesel meets the necessary performance and safety criteria for use in diesel engines. During the 1980s and 1990s, significant research and development (R&D) efforts were dedicated to biodiesel and its potential as a renewable energy source and substitute for conventional diesel fuel. This period was critical for establishing the technical foundation for biodiesel production, performance, and commercialization.Non-edible oils acquired from plant group such as Jatropha, Pongamia pinnata and others are the major commodity sources of biodiesel in India. Bio-diesel can be intermingled with petroleum diesel at any ratio to produce a bio-diesel blend or used in its pure form also.

Biodiesel shares many similarities with petroleum diesel and can operate in compression ignition (CI) engines with modest or no modifications. This is one of the key advantages of biodiesel as an alternative fuel for diesel engines. The similarities in properties make biodiesel a viable substitute for petroleum diesel in various applications, such as transportation, agriculture, and industry. Biodiesel has the advantage of being compatible with the existing fuel storage infrastructure used for petroleum diesel, which means that it does not require separate storage facilities or significant changes to infrastructure. This ease of integration into current systems makes biodiesel a convenient and cost-effective alternative fuel for diesel engines Nitrogen oxide (NOx) emissions tend to be higher when using biodiesel, particularly in higher concentrations (like B100 or 100% biodiesel). NOx is a pollutant that contributes to the formation of ground-level ozone, acid rain, and smog. The reason for the increase in NOx is that biodiesel generally has a higher cetane number (which improves ignition quality), leading to higher combustion temperatures.

These higher temperatures can promote the formation of NOx. Solution (EGR System): One potential way to mitigate this issue is through the integration of an Exhaust Gas Recirculation (EGR) system. EGR reduces NOx emissions by recirculating a portion of the exhaust gases back into the engine's combustion chamber, which helps lower the peak combustion temperature and, in turn, reduces NOx formation. Biodiesel has virtually no sulfur content, unlike petroleum diesel, which can contribute to the formation of sulfur dioxide (SO2) and acid rain. This makes biodiesel cleaner in terms of sulfur emissions. Biodiesel contains no aromatics, which are compounds typically found in petroleum diesel that can contribute to the formation of soot and particulate matter during combustion. This characteristic reduces the particulate matter emissions in biodiesel use. Biodiesel contains about 10% oxygen by weight, which allows it to burn more completely and efficiently compared to petroleum diesel. This in-built oxygen helps reduce the formation of carbon monoxide (CO) and unburned hydrocarbons (HC), which are common pollutants produced by incomplete combustion.

Several practices for manufacturing bio-diesel have been established. Stirring by Mechanical means, Cavitation method by Ultrasonic and hydrodynamic and super-critical methanol are a some samples. Some of the benefits of bio-diesel over fossil fuels like diesel and petrol are: Bio-diesel is a virtuous lubricant which is 65% better than petrol & diesel, produces a smaller amount smoke and particulate matter as Bio-diesel is sulphur free and aromatics. It has high cetane number, produces a smaller amount carbon monoxide and hydrocarbon releases, they are renewable, bio-degradable, and non-toxic. The key purposes of this work are the study of comparative of Gongura Seed bio-diesels and its mixture with Diesel in comparison to clean fuel (Diesel). The engine parameters to be evaluated and compared are (i) Brake Thermal Efficiency (BTE), (ii) Smoke Opacity, (iii) Brake Specific Fuel Consumption (BSFC) (iv) Carbon Monoxide (CO) emission (v) Hydrocarbon (HC) emission (vi) Nitrogen Oxide (NOX) release & (vii) Cylinder pressure. [1] conducted in three operating modes: diesel mode; methanol/diesel mode (methanol mode for short) and methanol-based SiO2 nanofluid (MSN)/diesel mode (MSN mode for short) where the MSN was injected in the intake manifold.

Experimental studies by [2] used Al2O3-based NF and Pg procedure, which resulted in higher levels of nitric oxide (NOx) by 32.6%, lower smoke emissions by 23.2%, lower hydrocarbon (HC) and carbon monoxide (CO) emission levels by 18.2-21.4%, and greater brake thermal efficiency by 11.5%. [3] said that the trans-esterification process was affected by a number of parameters, including the ratio of (Molar) oil and alcohol (1 vs 4), the quantity of reagent (catalyst – 0.1g), response time for accomplishment (1 Hour), and the process temperature (80 °C). Under the best conditions, the maximum rate of bio-diesel production was more than 92% at 80 °C for 1 hour reaction time. Presence of LaTiO3 NPs were recorded. [4] demonstrated the preparation of Citrus medica biodiesel by using green and eco-friendly nano-particles (NPs) of synthesized copper oxide. With using aqueous Portulaca oleracea extract. [5] conducted experiments with Six Set of Dual Biodiesel blends with four different Biodiesel. He concluded that at 4.4 HP Brake Power, the mix of D90 + NB5+ JB5 reduces the carbon monoxide by around 47% and BTE increased by about 37% for the blended fuel D90 + JB5+KB5. [6] Author conducted experiments with dual biodiesel with various concentration of Water v/v %.

Finally concluded that 5% of water is the best emulsified fuel. [7] He Conducted experiments with dual biodiesel blends (Rapeseed and Mahua) and concluded that for the blended fuel BL20 emission properties are reduced in comparison to Pure diesel. Slight increase in Nox level was witnessed. [8] the author found the performance and discharge character of several dual bio-diesel mixes were discovered in this article (mixture of jatropha bio-diesel and neem bio-diesel). The outcomes displays that sensible emission features for the blend (BB10 & BB20). [9] Findings the Properties of palm and castor biodiesel and conducted experiments of dual biodiesel blends ranging from 0 to 40%. [10] while his analysis increase in BTE and reduction in SFC was observed with lower value of biodiesel, he also found that emission levels were found lower than clear diesel. [11] have carried out various experiments with Dual Injection diesel engine with blends of palm methyl esters of different ratio. Tremendous change in performance were inked. [12] conducted number of experiments using Jatropha bio diesel with methanol as additives. His results concluded a good BTE for jatropha bio diesel, diesel and dual fuel with an ration of 29%, 30.2% and 28.7% respectively. [13] have done experiments with blends of two fuels pongamia, jatropha added with neat diesel. His results were blends of 20% & 10% of biodiesel with diesel produced near clear diesel performance and emission and suggested as substitute fuel for fossil fuel like diesel and petrol.

[14] investigatied on single cylinder diesel engine fuelled with cotton seed biodiesel blends with nano additives [15] found that a 20% blend of vegetable oil (such as Pongamia or Mustard oil) and 80% clean diesel could be used as an alternative fuel for diesel engines without requiring any modifications to the engine. This is significant because it demonstrates the engine compatibility of biodiesel-like mixtures. This blend showed lower emissions of carbon monoxide (CO) and hydrocarbons (HC) compared to conventional diesel, which is consistent with other biodiesel studies showing cleaner combustion and reduced particulate matter. The results indicate that such a blend can effectively reduce harmful pollutants while maintaining the operational capabilities of the engine. The research revealed that using the blend of vegetable oils and diesel did not result in significant power loss when compared to conventional diesel. This is crucial because the performance of the engine is one of the most important factors for adopting alternative fuels, and maintaining comparable or even improved performance with biodiesel blends makes it a viable option for many diesel engine applications. [16] has piloted tests on a single cylinder, oil cooled, four strokes, direct injection VCR diesel engine with different ratios of palm and jatropha bio-diesel with orthodox diesel fuels. Lower viscosity and better calorific values of bio-diesel fuel leads to greater power brake and vice versa was shown on his study. He has determined that the lower mixes of dual biodiesel blends on bio-diesel can be used as alternate fuel in diesel engine without any adjustment of engine. A.M.

[17] has studied the performance features & properties of dual bio-diesel mined from pongamia pinnate & neem oil. Author has concluded that the neem oil and pongamia oil is mono ester and is formed by means of trans-esterification process and it can be utilized in engine without any changes. [18-21] says that to change diesel fuel by bio-diesel, he has selected oils from palm kernel and eucalyptus. He added with methyl ester of palm kernel in the ratio of 15%, 10% and 5% by volume. He piloted a test in VCR diesel engine with mixed diesel fuel[22]. the incorporation of graphene oxide nano fuels into Tomato Oil methyl ester blends demonstrates potential for improving engine performance and reducing emissions[23]. he ratio of B20, B40, and B60 with Tio2 as a Nanoparticles as 100ppm. Performances and emission characteristics of engine is calculated with blended fuel and neat diesel[24-25].

II. METHODOLOGY

II.1 RESEARCH METHODOLOGY

Currently human all over the world is fronting with two chief tasks which include the fuel devastations and ecosphere dilapidation. With current utilizing rate the reserve of gasoline products is limited to next 40 years only and the cost of crude oil depends on international volatile markets. Numerous countries spread over the world have been evolving novel crops from 1975 in order to advance the bio-mass reserve for production of various forms of bio energy. India's dependence on imported petroleum is a significant issue for its economic stability. Currently, the country imports about 70% of its petroleum needs, making it highly vulnerable to fluctuations in the global oil market. These fluctuations can be driven by geopolitical tensions, supply disruptions, and other factors that lead to volatile oil prices. This reliance on imported oil puts a strain on India's foreign exchange reserves and affects its balance of payments, which is a key concern for the country's economic health. Properties of Gongura Seed Biodiesel and Diesel as given in Table 1& Table2. Research Methodology of the project as shown in figure 1.

Table 1: Properties of Gongura Seed Biodiesel.

Appearance	Light Yellow Liquid
Density	0.88 at 200C
Kinematic Viscosity	1.9-6.0 mm3/s at 400C
FFA Content	6.9 wt%
Iodine No	1.38 g12/100g-oil
Water Content	0.02 wt%

Source: Authors, (2026).

Table 2: Properties of Diesel.

Calorific Value	43200 kJ/kg
Density	830 at 150C
Kinematic Viscosity	3.1 at 400C
Flash Point	680C
Cetane Number	46.4

Source: Authors, (2026).

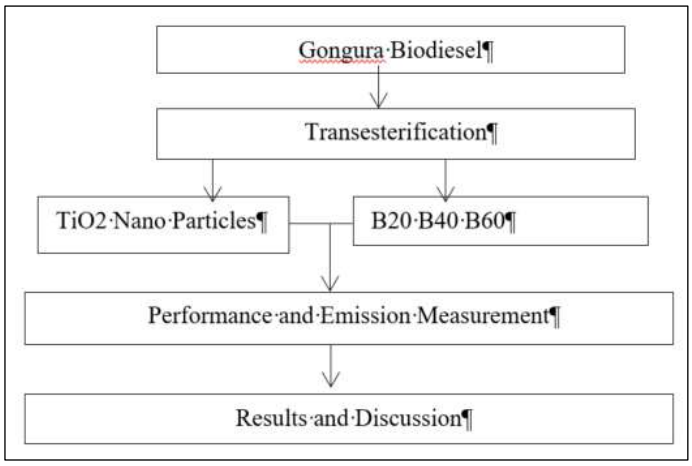


Figure 1: Research Methodology of the Project.
Source: Authors, (2026).

II.2 EXPERIMENTAL DETAILS

From the various ratio of mixed Gongura Seed biodiesel experiments is evaluated for B20, B40 & B-60 to find performances and emissions characteristics of the engine without any modification made on it and keeping the compression ratio as 17.5 running at constant speed 1500rpm. Engine Specifications as given in Table 3. Experimental Setup as shown in figure 2.



Figure 2: Experimental Setup.
Source: Authors, (2026).

Table 3: Engine Specifications.

Manufacturer	Kirloskar,AV-1,water Cooled 4 Stroke
Type	Direct Diesel Engine
Fuel	Diesel
Bore ,Stroke	87.5 x 110 mm
Compression Ratio	17.5
Speed Rpm	1500
Nozzle Opening Pressure	200
Rated Power	5.2 kw@ 1500 rpm

Source: Authors, (2026).

III. RESULTS AND DISCUSSION

III.1 CARBON MONOXIDE (CO)

The Emission of Carbon monoxide with different proportional of load for Gongura Seed biodiesel mix of 0%,20%, 40% & 60% with diesel and plain diesel were depicted in Figure 3. It was experiential during the test, lesser emission of CO was observed with increase in engine load during part load and increase in level during full load. It might be due the maximum fuel consumption due to rich air-fuel mixture. From the below figure we can come to a conclusion that throughout the engine load, substantial drop in CO emission of the Gongura Seed bio diesel blends in comparison to neat Diesel. It is primarily because of higher oxygen level in biodiesel than neat diesel.

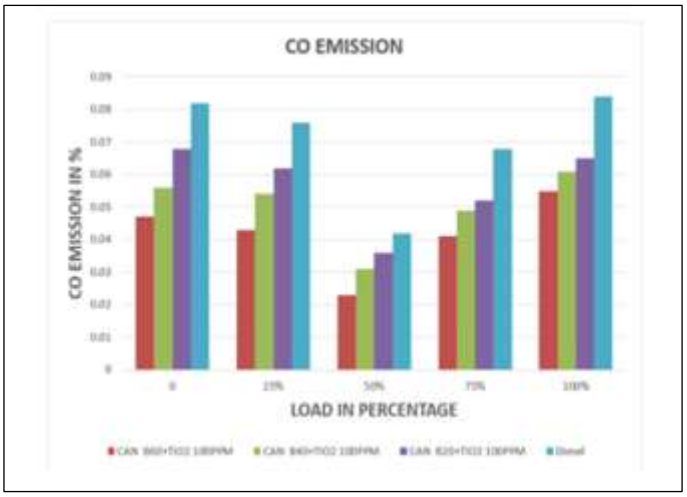


Figure 3: Load Vs CO Emission.
Source: Authors, (2026).

III.2 NITROUS OXIDE

Nox emission of neat diesel fuel and blends of Gongura Seed bio diesel is available from Figure 4. It was discovered that release of Nox was higher than pure diesel fuel for all Gongura Seed bio-diesel. The increase in combustion temperature and the increase in flame temperature could be key reasons for the higher NOx (Nitrogen Oxides) emissions observed when using biodiesel or vegetable oil blends in diesel engines. Nox formation was mostly due to surge in cylinder combustion temperature and percentage of oxygen level. From test results it was found that Nox release is more compared to pure diesel oil than mixes of Gongura Seed oil.

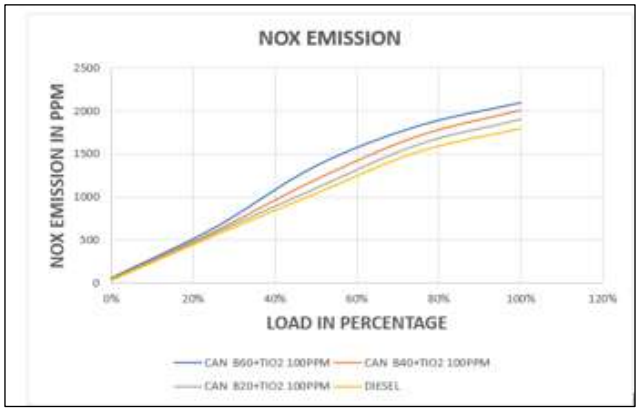


Figure 4: Load Vs NOX Emission.
Source: Authors, (2026).

III.3 HYDROCARBON

Differences of HC emission with engine load while using various mixes of Gongura Seed bio-diesel and pure diesel is shown in the Figure 5. During testing, for all the loads tested HC emissions were lesser at engine load although it has increased during higher engine load. This is mainly because of larger fuel particle size, timing of injection and chocking of nozzle also changed the combustion timing. It is mainly because of lesser oxygen presented during more fuel injection at higher loads. The key cause of un-burn Hydro carbon is owing to minimum ignition delay linked with fuels having considerable cetane number which reduces the over varied fuel. The above result settled with the references.

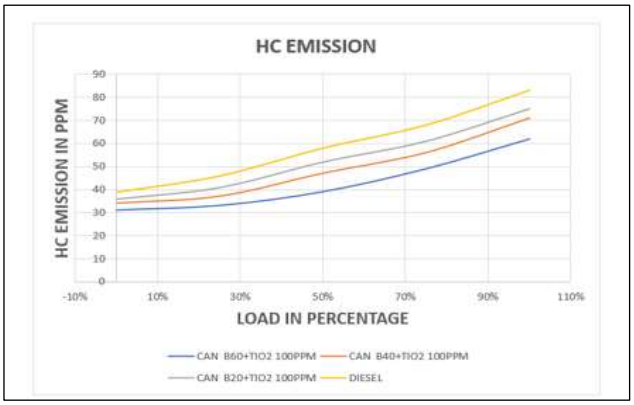


Figure 5: Load Vs HC Emission.
Source: Authors, (2026).

III.4 CO2 EMISSION

As Figure 6 represents the CO_2 emission of Gongura Seed biodiesel at 20%, 40% & 60% with pure diesel at different engine loads. During the increase in engine load, increase in CO_2 release was observed which might be due to greater fuel consumption linked with load raises. Lesser CO_2 release were observed for various mix of bio-diesel Gongura Seed and compared with clean diesel fuel. The drop in CO_2 release were clarified on relations of greater oxygen content in bio-diesel mixes of Gongura Seed as compared to pure diesel fuel.

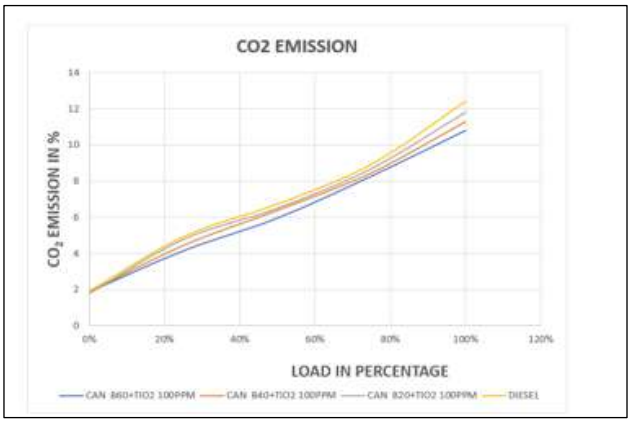


Figure 6: Load Vs CO2 Emission.
Source: Authors, (2026).

III.5 SMOKE OPACITY

The smoke opacity of all candle nut bio-diesel mixture with various engine load and pure diesel is shown in Figure 7. For comparison purpose clean diesel smoke opacity is also added in the graph. Similar to other experiments the result illustrates rise in smoke release with upsurge in engine capacity. It is primarily owing to change in fuel usage in the higher side which causes rich air-fuel mixture. As the percentage of oxygen molecules is more and carbon content is small in the Gongura Seed bio-diesel in comparison to pure diesel which has caused great burning and reduction in smoke emission was observed.

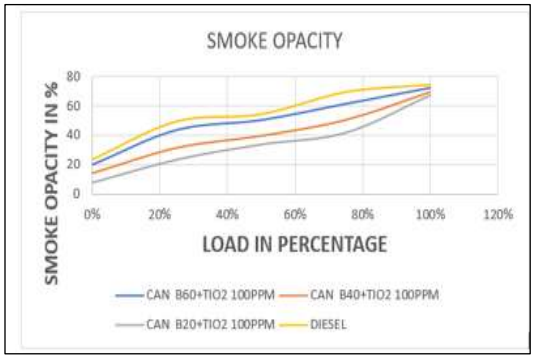


Figure 7: Load vs Smoke Opacity.
Source: Authors, (2026).

III.6 SPECIFIC FUEL CONSUMPTION (SFC)

The smoke Specific fuel consumption of Gongura Seed bio-diesel with various loads in the engine for different mixes were illustrated in Figure 8. Specific fuel consumption was found to be less with increase in percentage of load for all ratio of blended bio-diesel. It is mainly because of the kinematic viscosity and the good calorific values of the bio-diesel. Illustrations from all graphs shows clearly consumption of fuel is maximum because of the lower content of ester in the fuel.

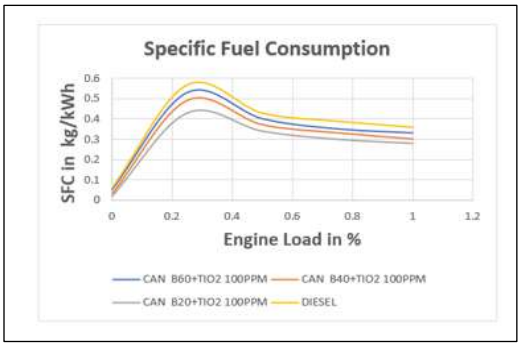


Figure 8: Load vs Specific Fuel Consumption.
Source: Authors, (2026).

III.7 BRAKE THERMAL EFFICIENCY (BTE)

From figure 9, we can understood that the brake thermal efficiency rises during increase in break power for diesel and candle nut biodiesel mixes. The brake thermal efficiency of diesel is higher equated to Gongura Seed bio-diesel blends. Also from the information, it is well-known that the BTE of lower ration blend is closer to that of pure diesel when compared to Gongura Seed bio-diesel blends. This disparity in BTE for various bio-diesel mixes is owing to higher viscidness and reduced instability, which principals to deprived blend formation. This results in reduction of brake thermal efficiency for Gongura Seed bio-diesel mixes.

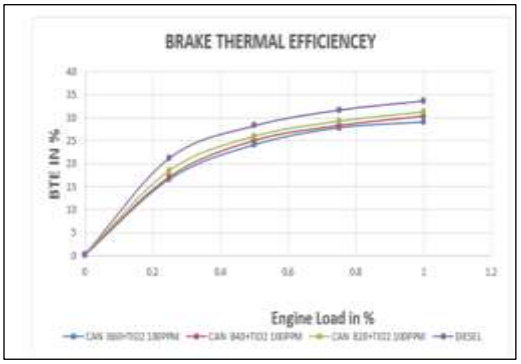


Figure 9: Load vs Brake Thermal Efficiency.
Source: Authors, (2026).

III.8 MAX CYLINDER PRESSURE (BAR)

The distinction of engine's cylinder pressure in response to various eccentric angle at various proportions of Gongura Seed Bio-diesel of various proportions and clean diesel at maximum load. It was witnessed from the experiment that maximum pressure of cylinder of Gongura Seed Bio Diesel mixes is lesser than that of pure diesel. It is chiefly due to higher viscosity and smaller volatility of Gongura Seed bio-diesel. From Figure 10, we can observe that the maximum engine cylinder pressure of diesel is 66 kg/cm2 and B20 is 59 kg/cm2. From the graph it is easily understood that the maximum engine cylinder is extra for diesel compare to Gongura Seed Bio-diesel mixes and reduces when bio-diesel of larger proportions of diesel was used. It is chiefly because of the delay in ignition time and lesser air/diesel or bio-diesel mixture which is accessible during the combustion & ignition. It has therefore added burning in the diffusion-burning stage than that of the pre-mixed burning point. Because of the above reasons the maximum engine cylinder pressure is smaller for Gongura Seed bio-diesel-diesel mixes.

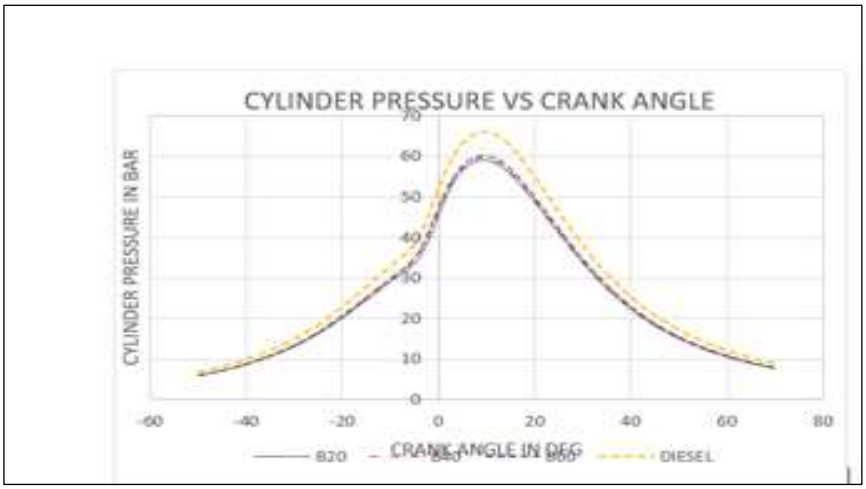


Figure 10: Load vs Cylinder Pressure.
Source: Authors, (2026).

IV. CONCLUSIONS

During the study a direct injection diesel cylinder with single cylinder was used to run with different blends of Gongura Seed biodiesel in the ratio of 20%, 40% & 60% with clean diesel. Performance study of engine and exhaust discharge characters were calculated during several engine load of 0%, 25%, 50%, 75% & 100% with engine speed of 1500 rpm. Emissions like CO, HC, smoke opacity, Co2 & No2 were scrutinized. Brake thermal efficiency, Cylinder Pressure and specific fuel consumption were also studied and related with pure diesel. Conclusions that have been found are briefed here as:

❖ Exhaust emission of carbon monoxide, smoke and hydro carbons are lesser for the experienced Gongura Seed bio-diesel blends B20, B40 & B60 when compared to clean diesel fuel.

❖ Carbon dioxide release through exhaust when using Gongura Seed bio-diesel of 20%, 40% & 60% blends were lesser compared to pure diesel as fuel.

❖ Release of NOx while using Gongura Seed bio-diesel of 20%, 40% & 60% blends improved when equated with pure diesel as fuel for the examined bio-diesel mixes.

❖ SFC for Gongura Seed biodiesel blends B20, B40 & B60 are lower in relation to diesel fuel and BTE of Gongura Seed bio-diesel mixes have approached the near value of clean Diesel.

❖ Gongura Seed Blends of B20 & B40 are having lower emissions and near BTE to diesel and can be used as substitute fuel.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Sathiyaraj.S, Kumaran.P, Imthiyas.A
Methodology: Sathiyaraj.S, Silamburajan.M, Rengaraju.G
Investigation: Sathiyaraj.S, Imthiyas.A
Discussion of results: Sathiyaraj.S, Imthiyas.A
Writing – Original Draft: Sathiyaraj.S, Kumaran.P, Imthiyas.A
Writing – Review and Editing: Kumaran.P, Imthiyas.A
Resources: Sathiyaraj.S, Silamburajan.M, Rengaraju.G
Supervision: Kumaran.P
Approval of the final text: Kumaran.P ,Imthiyas.A

VI. REFERENCES

[1]. Jiangjun Wei, Chengjun He, Gang Lv, Yuan Zhuang, Yejian Qian, Suozhu Pan,(2021), The combustion, performance and emissions investigation of a dual-fuel diesel engine using silicon dioxide nanoparticle additives to methanol, Energy, Volume 230,2021,120734,ISSN 0360- 5442. https://doi.org/10.1016/j.energy.120734.

[2]. Sateesh, K.A., Yaliwal, V.S., Soudagar, (2022). Utilization of biodiesel/Al2O3 nanoparticles for combustion behaviour enhancement of a diesel engine operated on dual fuel mode. J Therm Anal Calorim 147, 5897–5911.

[3]. ShahabaldinRezania, SaeidehMahdinia, Bahareh Oryani, Jinwoo Cho, Eilhann E Kwon, Alireza Bozorgian, Hamid Rashidi Nodeh, NegisaDarajeh, KamyarMehranzamir,(2022)" Biodiesel production from wild mustard (Sinapis Arvensis) seed oil using a novel heterogeneous catalyst ofLaTiO3nanoparticles,Fue" ,Volume,307.

[4]. Rozina, Shir Reen Chia, Mushtaq Ahmad, Shazia Sultana, Muhammad Zafar, Saira Asif, Awais Bokhari, Saifuddin Nomanbhay, Muhammad Mubashir, KuanShiong Khoo, Pau Loke Show, (2022) "Green synthesis of biodiesel from Citrus medica seed oil using green nanoparticles of copper oxide,Fuel",Volume 323,2022,124285,ISSN 0016-2361.

[5]. Siraj Sayyed, Randip Kumar Das, Kishor Kulkarni,(2022) "Experimental investigation for evaluating the performance and emission characteristics of DICI engine fueled with dual biodiesel-diesel blends of Jatropha" Karanja, Mahua,andNeem,Energy,Volume238,,121787,ISSN0360-5442.

[6]. Arnab Roy, YuvrajsinhDabhi, Hardik Brahmbhatt, Sajan K. Chourasia,(2021) "Effect of emulsified fuel based on dual blend of Castor-Jatropha biodiesel on CI engine performance and emissions" Alexandria Engineering Journal, Volume 60, Issue 1,2021,Pages 1981-1990,ISSN 1110- 0168.

[7]. A. Saravanan, M. Murugan, M. Sreenivasa Reddy, SatyajeetParida, (2020)" Performance and emission characteristics of variable compression ratio CI engine fueled with dual biodiesel blends of Rapeseed and Mahua, Fuel" Volume 263, 2020, 116751.

[8]. K. Arun Balasubramanian (2012) "Dual Biodiesel Blends in Diesel Engine - Performance and Emission Analysis" European Journal of Scientific Research SSN 1450-216X Vol.75 No.3 (2012), pp. 400-408.

[9]. Pala Tejesh and Vinod Kotebavi,(2019)" A comparative study on performance and emission characteristics of palm and castor dual biodiesel blends in CI engine" AIP Conference Proceedings 2200, 020078.

[10]. Pankaj Dubey, Rajesh Gupta, (2018)," Influences of dual bio-fuel (Jatropha biodiesel and turpentine oil) on single cylinder variable compression ratio diesel engine" Renewable Energy, Volume 115, 2018, Pages 1294-1302.

[11]. G. Baskar, G. Kalavathy, R. Aiswarya, I. Abarnaebenezer Selvakumari,(2019) " - Advances in bio-oil extraction from nonedible oil seeds and algal biomass, In Woodhead Publishing Series in Energy" Advances in Eco-Fuels for a Sustainable Environment, Woodhead Publishing,2019,Pages 187-210.

[12]. P. T. Saravankumar, V. Suresh, V. Vijayan & A. Godwin Antony (2019): Ecological effect of corn oil biofuel with SiO2 nano-additives, Energy Sources, Part A: Recovery, Utilization, and Environmental Effectshttps://doi.org/10.1080/15567036.2019.1576079.

[13]. Sanjay K. C., Shreyas, Vinol Pinto, Sheikh Shoaib Abdul Gafoor, Thilak Biju, K. Raju,(2016) "Effect of Alumina Nanoparticles on the Performance and Emission Characteristics of CI Engine Fuelled with Lard Oil Methyl Ester Blend", American Journal of Materials Science , 6(4A): 94-98.

[14].]Pallavi Ghogare, Dr. N. W. Kale,(2016) "Experimental Investigation on Single Cylinder Diesel Engine FuelledWith Cotton Seed Biodiesel Blends With Nano Additives", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016, ISSN: 2348 – 8360.

[15]. RolvinD'Silva, BinuK.G ,Thirumaleshwara Bhat,(2015) "Performance and Emission characteristics of a C.I. Engine fuelled with diesel and TiO2 nanoparticles as fuel additive" , Science Direct, Materials Today Proceedings 2,Pages:3728 – 3735.

[16]. M.Santhanamuthu, S.Chittibabu, T.Tamizharasan3 and T.P.Mani4,(2014) "Evaluation of CI engine performance fuelled by Diesel-Polanga oil blends doped with iron oxide nanoparticles" International Journal of ChemTech Research, ISSN : 0974-4290, Vol.6, No.2, pp 1299-1308, April-June 2014.

[17]. R. Sathiyamoorthi, M. Puviyarasan, B. Bhuvanesh Kumar And D. Breslin Joshua,(2016) "Effect Of CeO2 Nano Additive On Performance And Emission Characteristics Of Diesel EngineFuelled By Neem Oil-Biodiesel", Int. J. Chem. Sci.: 14(S2), , 473-484 ISSN 0972-768X.

[18]. Karthikeyan,(2014)" An environmental effect of GSO methyl ester with ZnO additive fuelled marine engine" Ind. J. Geo-Mar. Sci. 43 (4).
[19]. Ragu. R, Sai Krishnan. G, Dasaprakash. J., (2011)" Experimental Investigations On The Effects Of Aluminium Oxide Nanoparticles On Biodiesel" Advances in Natural and Applied Sciences. 11(9); Pages: 45-49.

[20]. Naveenap ,vinod R , prashanthreddy(2015) " Experimental Investigation On The Performance And Emission Characteristics Of Simarouda Glauca Oil Is An Alternate Fuel In A Variable Comptression Ignition Engine"international Journal Of Engineering Research And Technology , Vol 4 , Issue 06.

[21]. Swarup Kumar Nayak ,Bhabani Prasanna Pattanaik , (2014)," "Experimental Investigation On The Performance And Emission Characteristics Of A Diesel Engine Fuelled With Mahua Biodiesel Using Additive" Energy Procedia 54 ,Pg569-579.

[22]. K.Srithar , k. Arun balasubramanian ,A.Pavendan , B.Ashokkumar(2014) "Experimental Investigation On mixing of two biodiesels blended with diesel as alternative fuel for diesel" journal of king saud university Pg50-56.

[23]. Kumaran Pachaiyappan , Natarajan Sengodan , Prakash Sekar , Vasanthraj Ramalingam , Saranraj Periyasamy,(2024)" Enhancing Combustion Efficiency: Utilizing Graphene Oxide Nanofluids as Fuel Additives with Tomato Oil Methyl Ester in CI Engines" Turkish Journal of Engineering, Volume: 8 Issue: 4, 720 – 728.

[24]. K. Srithar , K. ArunBalasubramanian , V. Pavendan , B. Ashok Kumar(2017), "Experimental investigations on mixing of two biodiesels blended with diesel as alternative fuel for diesel engines" Journal of King Saud University – Engineering Sciences,, 29 Pg.No 50-56.

[25].Vidyasagar Reddy, GovindhaRasu(2016) "Experimental investigation on the performance of CI engine using dual biodiesel" 21st Century Energy Needs Materials, Systems and Applications (ICTFCEN), DOI: 10.1109/ICTFCEN.2016.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 244-258. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2877

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

PERFORMANCE AND EMISSION ENHANCEMENT OF A CRDI VCR ENGINE FUELED WITH TIO₂-INFUSED TOMATO SEED BIODIESEL USING RSM OPTIMIZATION

Kumaran P[1]*, Sathiyaraj S[2], Vijayakumar K[3], Albinraj[4] and Esakki mani K[5]

[1, 2, 3,4,5]Department Mechanical Engineering, Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research Foundation(deemed to be university), Salem, Tamilnadu, India.

[1]http://orcid.org/ 0000-0003-2781-8999, [2]http://orcid.org/0000-0001-8504-0412, [3]http://orcid.org/0000-0002-5964-982X
[4]http://orcid.org/0009-0003-7581-738X, [5]http://orcid.org/0009-0006-4366-9478

Email: *kumaranp@avit.ac.in, sathiyaraj@avit.ac.in, vijayakumar@avit.ac.in, albinjith94@gmail.com, shivasabari59@gmail.com

ARTICLE INFO

Article History
Received: October 31, 2025
Revised: November 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
CRDI engine,
Response surface methodology,
Optimization,
Alternative fuels,
Tomato seed biodiesel

ABSTRACT

This study enhanced the response of a CRDI single-cylinder four-stroke VCR engine under various input parameters with a novel tomato seed biodiesel (B25) infused with TiO2 nanoparticles. Response surface methodology (RSM) was employed to examine the influence of various input parameters on engine responses, including specific fuel consumption (SFC), brake thermal efficiency (BTE), and emissions of CO, HC, and NOx. The examined factors included injection pressure (IP), load, compression ratio (CR), and nanoparticle concentration as gasoline additives. Through the optimization of all input parameters, the RSM identified the optimal point at which the engine achieved maximum performance with minimal emissions. The optimal settings for the input parameters are as follows: a nano biofuel mix ratio, an injection pressure of 799 bar, a compression ratio of 20, and a load of 3.63 kg. This configuration of input parameters produces optimal outcomes for BTE (22.46%), SFC (0.25 kg/kWh), CO (0.023%), HC (22.01 ppm), and NOx (536 ppm). To ensure validity, we compared the experimental value with the ideal value predicted by RSM and observed a satisfactory correlation.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

1. INTRODUCTION

Biodiesel offers a renewable and cleaner alternative to conventional diesel, yet its performance in modern engines can be limited by factors such as fuel properties and combustion efficiency. Nanotechnology-based additives and advanced fuel systems have emerged as effective solutions to enhance engine output, fuel economy, and emission profiles. Common Rail Direct Injection (CRDI) technology enables precise control over injection parameters, facilitating optimal combustion. This study employs Response Surface Methodology (RSM) to systematically optimize performance and emission characteristics in diesel engines. Authors investigated the use of pulsed electric field (PEF) as a pre-extraction step to enhance TSO yield and quality. Optimum parameters were 9 min treatment and 0.5 µs pulse duration, resulting in maximum extraction of bioactives. Lycopene (565.54 mg/mL) was the main carotenoid, linoleic acid the predominant fatty acid; strong antioxidant activity contributed to high oxidative stability confirmed by EPR and DSC [1]. Authors studied supercritical CO_2 extraction (SFE) of lycopene from tomato pomace, analyzing lycopene concentration and E:Z isomer ratios. Initial process achieved 84.6% lycopene extraction but only 48.4% recovery in oil. Redesign using 703 g of exogenous tomato oil increased predicted lycopene recovery to 99.3%, showing that solubilization efficiency is driven largely by oil availability [2]. Review article summarizing extraction methods, physicochemical characteristics, and bioactive profile of TSO. Highlighted high levels of oleic and linoleic acids, potent antioxidant and antimicrobial properties, and applications in food products, packaging films, dietary supplements, and medical fields. Stressed potential for TSO to add economic value to tomato processing by-products [3].

Compared microwave-assisted extraction (MAE) and supercritical CO_2 extraction (SFE-CO_2) of TSO, optimizing via response surface methodology (RSM). MAE (56.2 °C, 29 min, 67.6 mL) produced higher yield (25.3 wt%) than SFE-CO_2 (16.9 wt%), with similar

fatty acid profiles (~80% unsaturated). γ-Tocopherol was dominant, linked to high DPPH• radical scavenging activity, suggesting strong potential for nutritional and food applications [4]. Researchers examined diesel–tomato seed oil biodiesel (TSOB) blends in a 4-cylinder, 4-stroke, indirect injection diesel engine using both experiments and AVL FIRE CFD simulation. Tested parameters included torque, BSFC, EGT, NOx, CO, CO_2, PM, in-cylinder pressure and temperature, and reaction progress at 1200–2400 rpm. B10 showed the lowest BSFC and B20 the highest torque. CFD revealed NOx formed mainly near the piston throat, CO/CO_2 near chamber boundaries, and PM at the injection region [5]. Researchers developed thiol-grafted Tamarindus indica seed gum–Teff hay biocarbon (TH@TI-TBC) adsorbent for cadmium removal from electroplating wastewater and achieved max adsorption (261.47 mg/g) at pH 5.5, following Langmuir and pseudo-second-order kinetics. Removed 89% of Cd from 30 mg/L wastewater; adsorption driven by electrostatic interaction and surface complexation. Authors proposed TH@TI-TBC as a sustainable waste-to-resource adsorbent [6]. Authors studied acetic acid–treated Pithecellobium dulce seed pod powder for Reactive Red 195-A dye removal. Characterized via SEM, FTIR, DTA, TGA; adsorption followed Temkin isotherm and pseudo-second-order kinetics ($R^2 > 0.90$).

Process was spontaneous and endothermic, suggesting PD pods as low-cost, eco-friendly adsorbents [7]. Researchers extracted 21.69% oil from Rio Grande tomato seeds via Soxhlet (79.83% unsaturated fatty acids), then recovered phenolics from defatted seeds via MAE optimized by RSM (700 W, 70 s, water). Combined oil and phenolic extraction added economic value to tomato waste, producing edible, biofuel, and cosmetic-grade oil and antioxidant-rich extracts through a green, low-cost process [8]. Authors studied turbocharger effects on a VCR diesel engine using neat diesel and neat TSO biodiesel. Turbocharging improved brake thermal efficiency and reduced CO and HC emissions for TSO compared to naturally aspirated operation, but NOx increased without turbo. With turbocharger, NOx emissions for TSO decreased relative to non-turbo TSO operation [9]. Researchers investigated nozzle hole number (2-hole vs 3-hole, 0.25 mm dia) on a single-cylinder DI diesel engine using diesel and B20 TSO at 1500 rpm, 200 bar injection pressure. The 3-hole nozzle improved brake thermal efficiency and reduced SFC at 40% load with B20, and produced lower NOx emissions than the 2-hole nozzle configuration [10]. The biodiesel combustion (apricot, papaya, sunflower, tomato seed oils) at compression ratios 15.5–18.5 in a Kirloskar AV1 diesel engine were simulated using ANSYS Fluent. Higher CR increased peak temperature, pressure, CO_2 and NO, and reduced CO

Apricot biodiesel at CR 18.5 achieved highest combustion efficiency but high emissions; tomato seed oil biodiesel at CR 15.5 showed lowest NO and CO_2 but reduce the performance [11]. Researchers reviewed tomato processing by-products (peels, seeds, pomace) as sources of bioactive compounds (carotenoids, proteins, fiber, pectin) and highlighted tomato seed oil's oxidative stability from carotenoid content and its applicability in food, feed, and biodiesel. The re-use of tomato residues in bakery products, animal feed, meat additives, and non-food industrial applications were analysed [12]. The biodiesel production from tomato seed oil (TSO) using Aspen Plus® were modeled, focusing on alternative catalysts and by-product recovery (glycerol, potassium phosphate). The methyl ester yield correlations and simulated transesterification with methanol recovery and recycle were developed. Designed distillation towers and calculated energy return on investment (EROI), including harvesting and processing energy. Results indicated TSO biodiesel production is technically and energetically feasible [13]. Diesel–TSOB blends were analyzed in an indirect injection (IDI) diesel engine through experiments and 2D CFD (AVL FIRE). CP and combustion characteristics were measured at five loads and seven speeds. B10 at full load had peak CP of 69 MPa at 2200 rpm shows highest accumulative heat release (~1900 J), and best fuel energy efficiency. CFD showed B5 had longest jet penetration (44 mm), while B10 reached peak in-cylinder temperature (1600 K) [14]. Ultrasound-assisted protein extraction from defatted tomato seed cake was optimized using RSM.

The optimum (52 mL/g, 2 min 18 s, 60 min) achieved 35.75% protein yield. The extracted protein exhibited favorable physicochemical properties suitable for food industry applications [15]. Cold press extraction (12.80% yield) and enzyme-assisted aqueous extraction (9.66% yield) for tomato seed oil was analyzed. Both eco-friendly methods preserved oil quality and bioactivity, highlighting their potential for functional food and nutraceutical applications [16]. A UiO-66-NH_2/ZnO/TiO_2 nanocatalyst via sol–gel for biodiesel production was developed from dairy waste scum oil (DWSO). It exhibited a high surface area (568.1 m^2/g) and small particle size (28.35 nm). Using CCD optimization, a 98.7% conversion was achieved at a 9.8:1 methanol/oil ratio, 61 °C, and 2 wt% catalyst [17]. Diesel engine performance was optimized using ternary blends of n-heptane, biodiesel, and diesel with TiO_2 nanoparticles (0–75 ppm) under varying loads, employing RSM. Optimum at 76.8% load, 57.6 ppm NPs, 25.6% biodiesel, and 9.5% n-heptane achieved 77.54 bar CP, 64.65 kJ HRR, 35.12% BTE, and reduced CO (0.0647% vol), HC (52.11 ppm), and smoke (37.34%). Validation errors were <5% [18]. Researchers compared the experimental and ANN-based predictions for CI engine parameters with mahua biodiesel (M30) and 50 ppm of CeO_2, CuO, or TiO_2 NPs. CeO_2 NP blend improved BTE by 9.1%, CP by 11.3%, HRR by 10.2%, and reduced BSFC by 13.7%, CO by 30.4%, HC by 30.1%, smoke by 34.7%, and NOx by 7.1%.

ANN backpropagation with two inputs (load, blend type) effectively predicted performance and emissions, enabling optimal blend selection without extensive testing [19]. Synthesized $ZnFe_2O_4$@TiO_2 nanocatalyst for biodiesel production from jojoba oil, achieving 97.56% yield at 54.96 °C, 2.89% catalyst, and 12.13 mol/mol methanol/oil ratio. Catalyst retained >90% yield after seven cycles. Engine tests showed reduced CO, NOx, and UHC, along with improved BTE and BSFC. Reaction was endothermic ($\Delta H° = 44.82$ kJ/mol) [20]. Biodiesel from a mix of Ceiba pentandra, Mahua longifolia, and Azadirachta indica oils were produced and enhanced with CaO–TiO_2 nano-additives. Optimization by Box–Behnken Design, Deep Belief Network (DBN), and hybrid Arithmetic Optimization was performed. Algorithm identified optimal conditions (load 60%, biodiesel 25%, nano-additive 20 wt%, 1400 rpm). Compared to diesel, CaO–TiO_2 biodiesel reduced BSFC by 41%, CO by 32%, NOx by 25%, and increased BTE by 16.9% [21]. The previous studies examined the individual variables such as biodiesel addition, nanoparticle concentration fluctuations, injection pressure, and compression ratios, additional research into their combined impact is necessary. The author discovered research focused on improving biodiesel volume for engines through the manipulation of compression ratio, load, injection pressure, and nanoparticle concentration.

Therefore, studies that separately control fuel process parameters while jointly managing engine parameters may gain from the present research. Consequently, RSM has been investigated in this study to enhance engine performance and emission metrics. The optimal parameters for operation include nanoparticle concentration, load, CR, and IP. This study involved the incorporation of TiO2 nanoparticles into biodiesel derived from tomato seeds. The engine compression ratio was altered from 18 to 20 in increments of 0.5, the injection pressure was increased from 600 bar to 800 bar in 50 bar increments, and the volume fraction of the biodiesel blend was set at 25%. The concentration of TiO2 nanoparticles was modified in increments of 20 ppm, reaching a maximum of 80 ppm. We employed 2-kilogram increments to adjust the load from 0 to 8 kilograms.

II. MATERIALS AND METHODS

II.1 BIOFUEL PREPARATION PROCESS

We procured the tomato seed oil from the local market in Chennai, Tamil Nadu, India. The subsequent step involved performing the fuel test. Fuel testing indicates that the oil's elevated viscosity renders it unsuitable for direct application in a diesel engine. The objective of the transesterification procedure was to reduce viscosity. The synthesis of biofuels using transesterification is a recognized technique. Subsequently, the mixture was exposed to temperatures between 60 and 70°C in the presence of a suitable catalyst of KOH (1% w/w). The flask was subjected to magnetic stirring at 400 rpm with the mixture. After one hour, a substantial layer had developed in the flask. Following a duration of two hours, the glycerol was isolated by the sedimentation process. B25 is a biofuel mixture including tomato seed oil. The Physio-chemical characteristics of the developed nanofuel is given in table 1.

Table 1: Physio-chemical characteristics of the developed nanofuel.

Oil	Kinematic viscosity (mm^2/s) at 40°C	Density (g/m^3)	High calorific value (kJ/kg)
Tomato seed oil	51.80	917	36,250
Supporting biodiesel B25 (TSO-25%	2.5120	0.8180	40,420
B25+20PPM(TiO_2)	2.6850	0.8235	40,570
B25+40PPM(TiO_2)	2.7105	0.8332	40,810
B25+60PPM(TiO_2)	2.7380	0.8431	41,030
B25+80PPM(TiO_2)	2.8995	0.8795	41,280

Source: Authors, (2026).

II.2 PREPARATION OF NANO-BIODIESEL BLEND

Table 2: Rroperties of TiO2 nanoparticle.

Appearance		Safety related information		Toxicological information	
Form	Fine powder	Boiling point	2972 °C	Inhalation	Avoid inhalation of dust; may cause mild respiratory irritation
		Melting point	1843 °C	Toxicity	Generally considered non-toxic, but nanoparticles may pose inhalation risk
Adour	Odorless	pH	Insoluble in water	-	-
Color	White	Flash Point	N/A	-	-

Source: Authors, (2026).

Table 2 presents the physicochemical parameters of the nanoparticles. The XRD technique is favored for the direct analysis of nanoparticle sample characteristics, including composition size, crystal structure, morphology, purity, and unit dimensions. The subsequent stage involved employing an ultrasonicator to disperse TiO2 nano powder in B25 at different concentrations (20, 40, 60, 80 ppm). Nanoparticles are incrementally added to B25 at consistent intervals for each sample to prevent micromolecule aggregation. Following one hour of high-frequency operation, the Ultrasonicator has completely dissolved all homogenous nanoparticles in the liquid. Consequently, four novel nano-biodiesel blends have been created, comprising 20 PPM (TiO2), 40 PPM (TiO2), 60 PPM (TiO2), and 80 PPM (TiO2) Titanium dioxide.

III. EXPERIMENTAL SET-UP AND METHODOLOGY

III.1 THE SPECIFICATION OF THE EXPERIMENTAL SETUP

This study utilized a CRDI VCR, a single-cylinder diesel engine. The schematic diagram is shown in figure 2. Table 3 presents the specification.

Table 3: CRDI VCR diesel engine specification.

Engine	Model	No. of cylinder	No. of stroke	Orifice diameter	Power	Speed	IP	CR range
Diesel	CRDI VCR	1	4	20 mm	3.5 kW	1500 rpm	60 MPa	12:1 to 18 :1

Source: Authors, (2026).

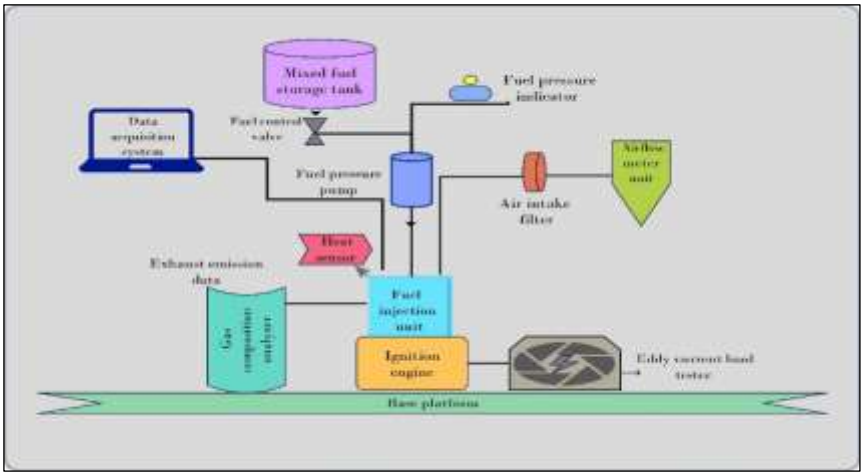


Figure 2: Block diagram of CRDI VCR engine.
Source: Authors, (2026).

III.2 RSM MODELING

III.2.1 Statistical analysis of response surface methodology

Table 4: Factors and Levels of input factors.

Input parameters	Coded levels					
	Symbol	-2	-1	0	1	2
A: Fuel	ppm	B25 (0)	B25 + 0.2 TiO_2 (20)	B25 + 0.4 TiO_2 (40)	B25 + 0.6 TiO_2 (60)	B25 + 0.8 TiO_2 (80)
B: LOAD	kg	0	2	4	6	8
C: IP	bar	600	650	700	750	800
D: CR	-	18	18.5	19	19.5	20

Source: Authors, (2026).

RSM is a data-centric methodology that enhances the output variable, influenced by several independent factors, by selecting the regression model based on quantitative data from the associated study. Response surface methodology is utilized to identify nonlinear interactions between inputs and outputs. Table 4 presents the optimal and non-optimal values of the input parameters corresponding to the -2, -1, Centre (0), +1, and +2 numerical values in this optimization study, as determined by statistical analysis. The objective of the central composite rotational design (CCRD) is to identify the optimal values for the input parameters and their interrelationships. Table 5 presents the experimental configuration for the CCRD design. While the Taguchi technique and factorial design are not ideally suited for optimizing nonlinear systems, they can do so effectively. Response surface methodology establishes a numerical relationship that encompasses the interactions between input parameters and output responses, which is challenging to achieve using optimization techniques; this constitutes the primary advantage of its application in the present study. In optimization approaches, it is essential to ascertain the relationship between input variables and output variables.

Table 5: Experimental array.

Run	A: Fuel (TiO_2 concentration)	B: LOAD	C: IP	D: CR	BTE (%)	SFC (kg/kWh)	CO (%)	HC (ppm)	NO_x (ppm)
1	60	6	750	19.5	30.95	0.432	0.0377	35.42	648
2	40	4	700	19	25.21	0.536	0.0501	27.93	388
3	40	4	700	19	25.4	0.541	0.0504	28.12	389
4	40	4	800	19	25.48	0.502	0.0373	24.56	512
5	60	2	750	18.5	19.95	0.6957	0.057	20.71	310
6	20	2	750	18.5	18.65	0.7507	0.0608	17.4	296
7	60	6	750	18.5	30.68	0.4317	0.0392	34.94	601
8	20	2	650	19.5	18.57	0.7611	0.0622	20.99	226
9	20	2	750	18.5	18.64	0.7507	0.0608	17.4	296
10	40	4	700	18	24.88	0.5478	0.0513	27.38	347
11	40	4	700	19	25.21	0.536	0.0501	27.93	388
12	40	4	600	19	24.24	0.568	0.0532	30.92	301
13	20	6	650	19.5	28.45	0.5261	0.0444	35.22	505
14	40	4	700	19	25.21	0.536	0.0501	27.93	388
15	40	4	700	18	25.54	0.5215	0.0489	28.64	437
16	80	6	650	18.5	30.37	0.4847	0.0424	38.24	485
17	20	6	750	19.5	28.75	0.4979	0.0415	31.99	612
18	0	4	700	19	22.07	0.601	0.0543	25.2	364
19	20	6	750	18.5	28.45	0.5107	0.0427	31.44	574
20	40	0	700	19	5.03	0.696	0.0672	14.36	163
21	40	8	700	19	32.24	0.511	0.0429	46.82	799
22	20	2	750	19.5	18.75	0.532	0.0587	17.62	341

23	40	4	700	19	25.21	0.536	0.0501	27.93	388
24	40	4	700	19	25.4	0.541	0.0504	28.12	389
25	60	2	650	19.5	19.48	0.7061	0.0584	24.42	240
26	60	6	650	19.5	30.75	0.4711	0.0412	38.75	533
27	60	2	750	19.5	20.45	0.6821	0.0558	21.38	358
28	20	6	650	18.5	28.07	0.5447	0.0462	34.74	457
29	40	4	700	19	25.21	0.536	0.0501	27.93	388
30	60	2	650	18.5	19.74	0.7347	0.0605	24.2	195
31	60	6	750	19.5	30.95	0.432	0.0377	35.42	648

Source: Authors, (2026).

The reduction of optimization trials is a subsequent advantage of RSM. Additional experimental trials are necessary for precise numerical validation of any approaches other than response surface methodology. The primary function of numerical models is to predict the behavior of input parameters and outcomes. The optimization method employs a higher-order polynomial equation to numerically model the nonlinear relationship between inputs and outputs. The subsequent equation denotes a second-order polynomial model [22].

$$Y = b_o + \sum b_i x_i + \sum b_{ij} x_i x_j (i, j = 1,2,3 \ldots \ldots k)$$

The variables Xi and Xj represent input variables, bi and bj denote constants, and Y signifies the output response. ANOVA was applied to analyze performance and emission parameters [23].

III.2.2 ANOVA results

Table 6: Analysis of variance for SFC and BTE.

Source	df	BTE		SFC	
		F-value	p-value	F-value	p-value
Model	10	16.91	< 0.0001	5.54	0.0007
A-Fuel	1	3.41	0.0805	0.5688	0.4600
B-Load	1	157.66	< 0.0001	44.33	< 0.0001
C-IP	1	0.4262	0.5217	2.09	0.1648
D-CR	1	0.0042	0.9491	0.2310	0.6363
AB	1	0.0333	0.8571	0.8454	0.3694
AC	1	0.0165	0.8990	0.1310	0.7214
AD	1	0.0172	0.8970	0.1535	0.6995
BC	1	0.0065	0.9365	0.0199	0.8893
BD	1	0.0164	0.8996	0.4177	0.5258
CD	1	0.0257	0.8744	1.49	0.2377
Residual	19				
Lack of Fit	11	240.96	< 0.0001	134.95	< 0.0001
Pure Error	8				
Cor Total	29				

Source: Authors, (2026).

Tables 6 and 7 present the findings of the analysis of variance (ANOVA) conducted on the experiments.

Table 7: Analysis of variance for $\boldsymbol{CO}, \boldsymbol{HC}$, and $\boldsymbol{NO_x}$ emission.

Source	DF	CO (%)		HC (ppm)		NO_x (ppm)	
		F-value	*p*-value	*F*-value	*p*-value	*F*-value	*p*-value
Model	10	17.41	< 0.0001	132.61	< 0.0001	51.38	< 0.0001
A-Fuel	1	6.86	0.0169	46.64	< 0.0001	2.75	0.1135
B-Load	1	151.15	< 0.0001	1098.49	< 0.0001	447.08	< 0.0001
C-IP	1	12.56	0.0022	49.01	< 0.0001	62.57	< 0.0001
D-CR	1	0.6844	0.4184	0.8237	0.3755	6.75	0.0176
AB	1	0.0059	0.9395	0.1518	0.7012	0.0895	0.7681
AC	1	0.1088	0.7452	0.3003	0.5901	0.0696	0.7948
AD	1	0.0001	0.9942	0.3942	0.5376	0.1028	0.7520
BC	1	0.1493	0.7035	0.1151	0.7381	0.0002	0.9901
BD	1	0.0005	0.9832	0.2271	0.6391	0.0084	0.9279
CD	1	0.0047	0.9461	0.0095	0.9232	0.0157	0.9016
Residual	19						
Lack of Fit	11	42.73	< 0.0001	17.79	0.0002	3.05	0.0622
Pure Error	8						
Cor Total	29						

Source: Authors, (2026).

A p-value below 0.01 signifies that the model is statistically significant. All models exhibit a p-value of less than 0.01. Consequently, ANOVA demonstrates that the model is statistically significant. The model F-values for SFC, BTE, CO, HC, and NOx are 390.07, 5.54, 16.91, 17.41, 132.61, and 51.38, respectively. The research indicated that nanoparticles and load substantially influence SFC (p-value <0.01). The suggested model and experimental variables can be evaluated using the P-value and F-value. Table 8 presents R2 values of 0.7445%, 0.8990%, 0.9016%, 0.9859%, and 0.9643% for SFC, BTE, CO, HC, and NOx, respectively, as indicated by the ANOVA results. The model's accuracy and adequacy are evidenced by the R2 value approaching 1. The R2 values indicate the efficacy of the models. The values of the Adjusted R2 (Adj. R2) and the Predicted R2 (Pred. R2) are relatively similar

Table 8: Model summary of input parameters.

Coefficient of determination	Value				
	BTE	SF	CO	HC	NO_x
R^2	0.8990	0.7445	0.9016	0.9859	0.9643
Adjusted R^2	0.8458	0.6100	0.8498	0.9784	0.9456
Predicted $\boldsymbol{R^2}$	0.7992	-0.0290	0.7910	0.9603	0.9340

Source: Authors, (2026).

IV. RESULT AND DISCUSSIONS

IV.1 ENGINE PERFORMANCE

IV.1.1 Specific fuel consumption (SFC)

Specific fuel consumption (SFC) refers to the amount of fuel burned per unit of time to provide a specified engine output. SFC is a quantitative measure for assessing an engine's efficacy in converting fuel into useable power. The initial equation is a second-order quadratic regression equation formulated to estimate SFC utilizing the coded values.

SFC = 0.573 -0.0200 A -0.1701B -0.0378 C -0.01248 D -0.0569 AB -0.02176 AC + 0.0232 AD -0.00908 BC + 0.0409 BD - 0.0798CD (1)

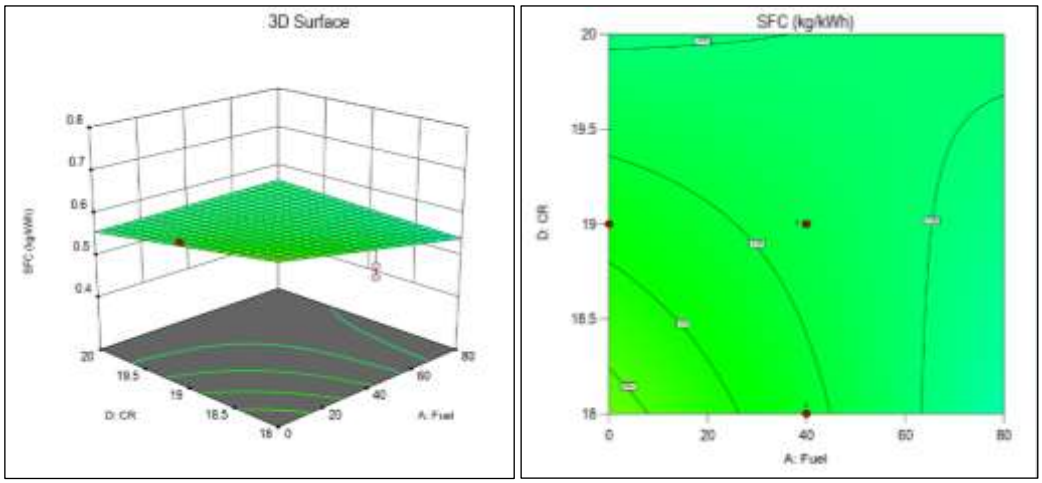


Figure 3: Three-dimensional plot depicts the impact of fuel and CR and their combined effect on the specific fuel consumption.
Source: Authors, (2026).

The influence of TiO_2 concentration in biodiesel and CR on SFC at IP (700 bar) and load (6 kg) was illustrated by contour and surface plots. At elevated CR, the light green hue of the contour zone yields a substantial concentration of nanoparticles (TiO_2) and a reduced engine specific fuel consumption (SFC). Biodiesel and CR exhibited a crimson hue with an escalating concentration of nanoparticles (TiO_2). As the TiO_2 to fuel ratio grew, the specific fuel consumption (SFC) value was found to decrease in figure 3. In another study, tested diesel–tomato seed oil biodiesel blends enhanced with nano-CuO (50 nm) additives in a single-cylinder engine. Nano-CuO improved fuel properties (flash/fire point, calorific value) and reduced SFC [24]. This is due to the fact that the expansion of nanoparticles improved the fuel's ignition characteristics by optimizing its air-fuel mixture, physicochemical properties (demonstrating increased fuel calorific value), dispersion, and thermal conductivity. A little reduction in SFC with increased CR indicates that enhanced combustion and diminished heat losses contributed to the overall decrease in SFC as CR increased. Under maximum load, the specific fuel consumption decreased by approximately 35% as the compression ratio increased from 18 to 20.

Figure 4 illustrates the effect of elevated injection pressure and compression ratio on specific fuel consumption (SFC) for a load of 4 kg and a fuel addition setting of 0. As IP increases, SFC diminishes with elevated CR. Decreased fuel usage results from increased spray pressure and a shorter duration for air/fuel combination. Figure 5, a contour plot depicts the influence of load and nano ratio on specific fuel usage. Figure 6 illustrates the trend of the specific fuel consumption curve in relation to the indicated power and fuel concentration, under a constant compression ratio and load. As the nanoparticle concentration and injection pressure increased, the contour plot indicates a drop in SFC. With a constant parameter load of 4 kg and CR, the SFC curve exhibits a slight increase or reduction contingent upon the IP level, as the nanoparticle concentration rises.

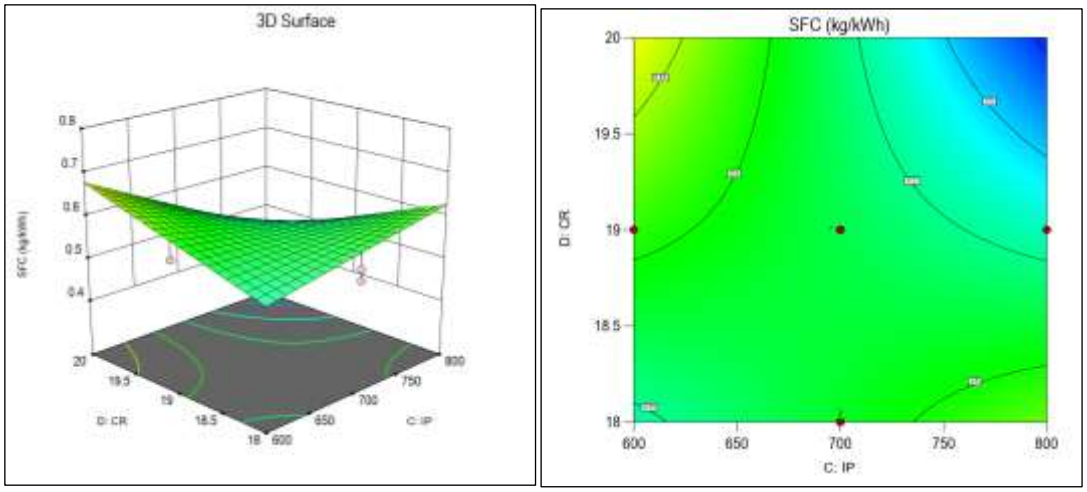


Figure 4: Three dimensional plot depicts the impact of IP and CR and their combined effect on the specific fuel consumption.
Source: Authors, (2026).

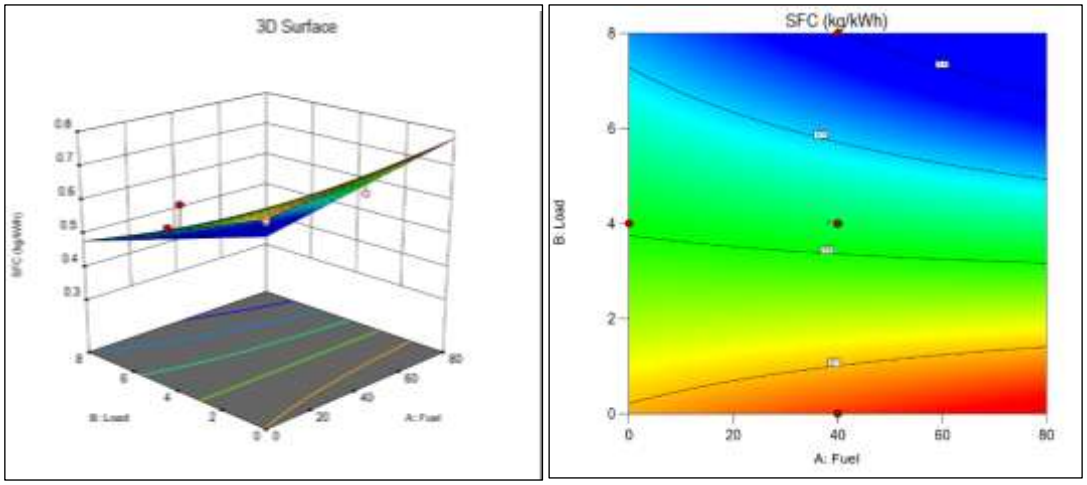


Figure 5: Three dimensional plot depicts the impact of fuel and load and their combined effect on the specific fuel consumption.
Source: Authors, (2026).

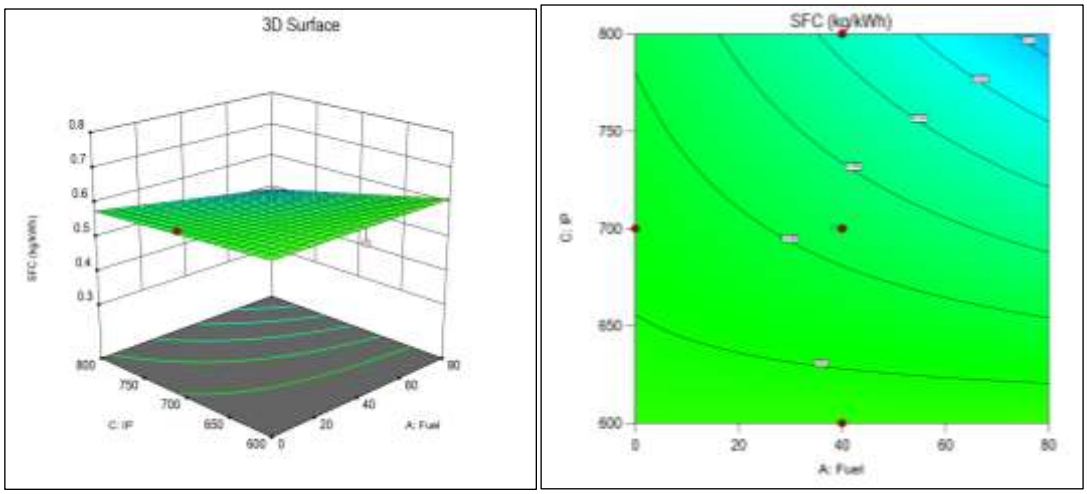


Figure 6: Three-dimensional plot depicts the impact of fuel and IP and their combined effect on the specific fuel consumption.
Source: Authors, (2026).

IV.1.2 Brake thermal efficiency (BTE)

To estimate BTE (utilizing the coded values), a second-order quadratic regression equation was formulated and denoted as equation (2).

BTE =24.23 + 1.744 A11 + .3809 B0 +.6056 C-0.0595 D0.+ 40084877135602 AB0.+.2742 AC0 +.27580 AD-0.1844 BC0.+ 28710 BD-0.3719 CD (2)

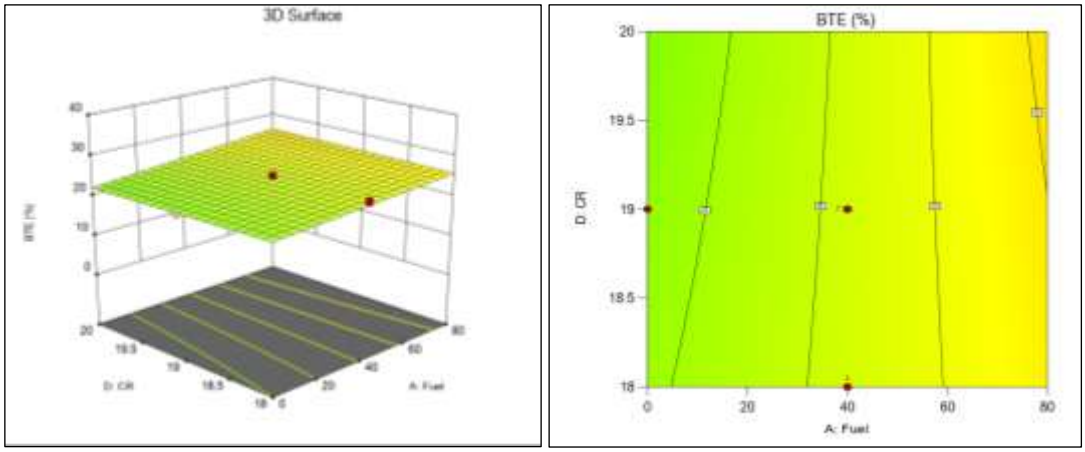


Figure 7: Three dimensional plot depicts the impact of fuel and CR and their combined impact on the brake thermal efficiency.
Source: Authors, (2026).

The Brake Thermal Efficiency (BTE) of the engine is a quantifiable indicator that assesses the engine's efficiency. Fuel conversion efficiency is the process of multiplying the lower heating value by the mass flow rate of injected fuel to produce energy. Using a load of 4 kg and an inlet pressure of 700 bar as fixed parameters, Figure 7 shows the contour and surface plot on the the influence of nanoparticle content and compression ratio on brake thermal efficiency. The addition of nanoparticles to biofuel resulted in an increased compression ratio and an elevated brake thermal efficiency curve. The utilization of nanoparticles signifies enhanced combustion and the effective transfer of fuel energy to productive labor. Optimal braking thermal efficiency was attained owing to the elevated heating value of the nano fuel and the substantial nanoparticle concentration within the fuel. An augmentation in BTE was seen with IP, as illustrated in Figure 8, demonstrating the impact of IP on BTE. Fuel injection transpires with reduced droplet sizes as the Fuel Injection Pressure (FIP) is elevated. This enhances the fuel's amalgamation with air and accelerates its vaporization. The combustion efficiency is improved, resulting in an increased brake thermal efficiency (BTE). In another study, lemongrass diesel engine tests showed B30ZnO had the highest BTE increase (>3%) and better emissions than TiO_2 or Al_2O_3 blends, with ANOVA confirming significant influence of nanoparticle type on performance [25].

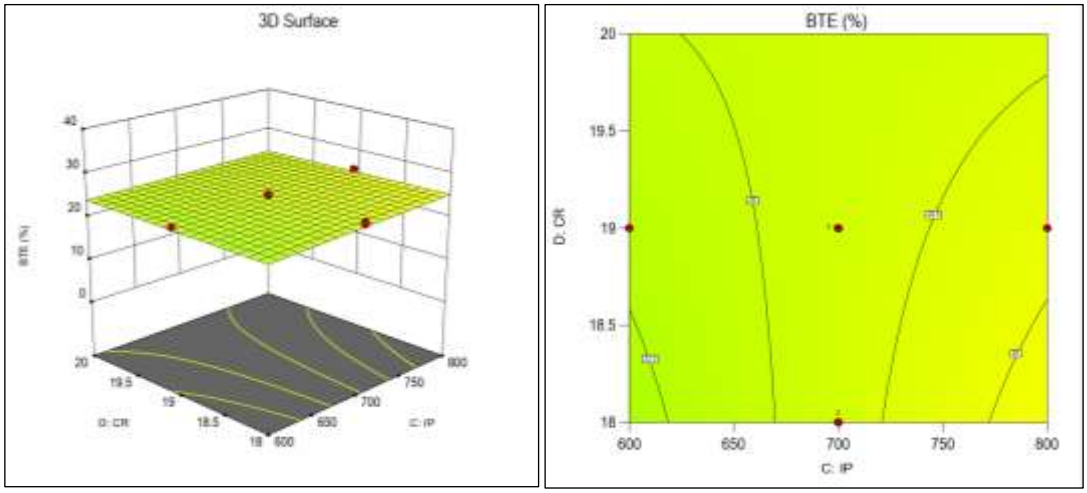


Figure 8: Three dimensional plot depicts the impact of IP and CR and their combined impact on the brake thermal efficiency.
Source: Authors, (2026).

Figure 9 illustrates the impact of load and nano fuel on BTE when the hold values are CR (19) and IP (700 bar). As the engine load intensified and the concentration of metallic nanoparticles augmented, its brake thermal efficiency enhanced. A minimal degree of heat loss is purported to enhance power as the load increases. Due to their elevated surface-to-volume ratio, nanoparticles facilitate the complete combustion of test fuel, hence enhancing brake thermal efficiency through nanoparticle advancement. Figure 10 illustrates the increase in BTE corresponding to the rising concentrations of IP and nanofuel. IP and nanoparticle enhancement improve fuel atomization, dispersion, and mixing, resulting in increased brake thermal efficiency.

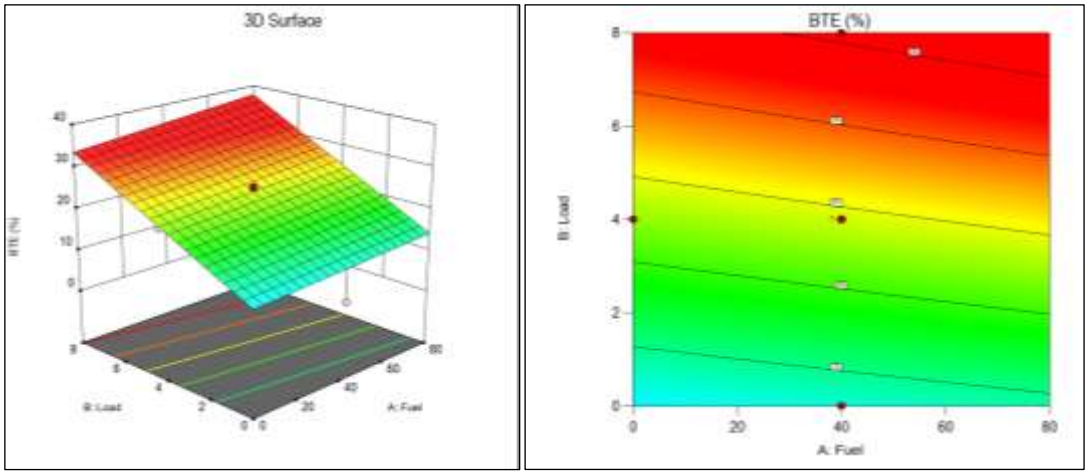


Figure 9: Three dimensional plot depicts the impact of fuel and load and their combined impact on the brake thermal efficiency.
Source: Authors, (2026).

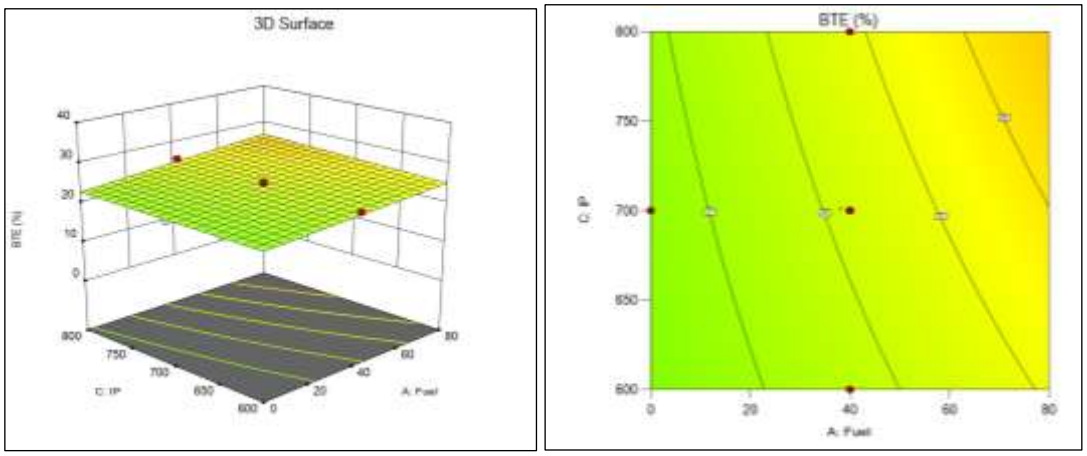


Figure 10: Three-dimensional plot depicts the impact of fuel and IP and their combined impact on the brake thermal efficiency.
Source: Authors, (2026).

IV.2 EMISSION CHARACTERISTICS

IV.2.1 CO emissions

The encoded values are utilized to formulate a second-order quadratic regression equation, employed to estimate CO as given in eqn (3).

$$CO = 0.05050 - 0.0035\,A - 0.0157\,B - 0.0046\,C - 0.0010\,D0 + .000239\,AB - 0.0009\,AC + 2.1932\text{-}05\,AD - 0.00125\,BC6 + 79151\text{-}05\,BD - 0.0002\,CD \quad (3)$$

Predictions for the response for specified levels of each element can be derived using the equation in terms of coded factors. The standard classification for the factors is +1 for elevated levels and -1 for diminished levels. To ascertain the relative significance of the factors, it is beneficial to compare their coefficients utilizing the coded equation.

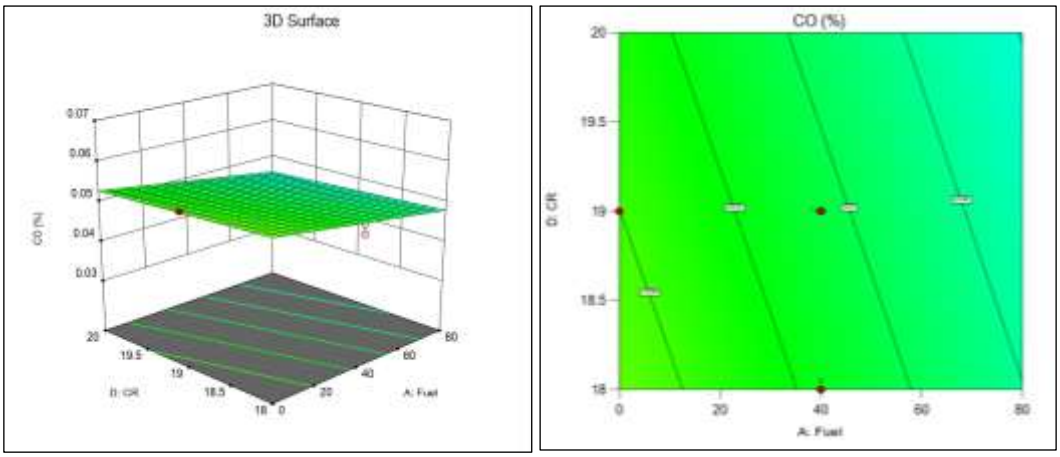


Figure 11: Three dimensional plot depicts the impact of fuel and CR and their combined impact on the CO emissions.
Source: Authors, (2026).

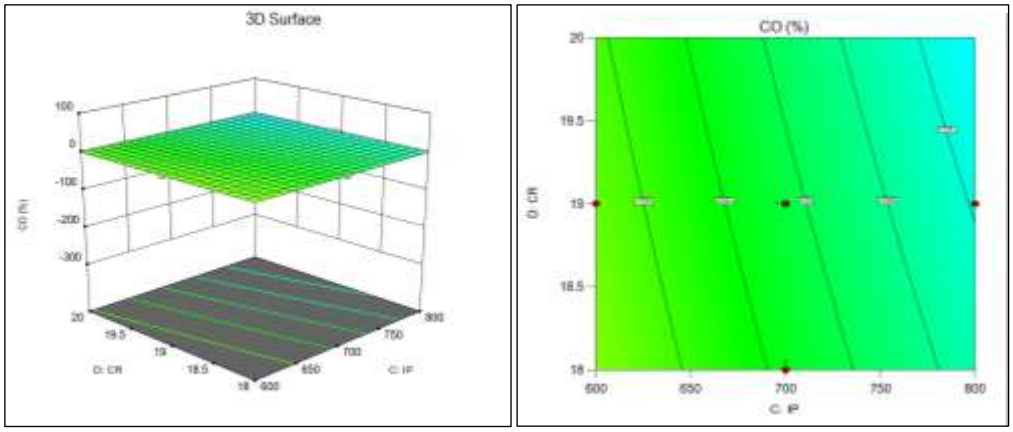


Figure 12: Three dimensional plot depicts the impact of IP and CR and their combined impact on the CO emissions.
Source: Authors, (2026).

The principal cause of carbon monoxide generation is the incomplete combustion of fuel. Figure 11 illustrates the changes in carbon monoxide for CR and nano fuel. The integration of nanoparticles reduced the carbon concentration due to complete burning. Metallic nanoparticles function as oxidation catalysts, transforming carbon monoxide into carbon dioxide during combustion. Metal nanoparticles also induced secondary atomization of the fuel. At elevated CR, the contour diagram indicated a decreased trend in CO emissions attributed to the augmented presence of nanoparticles. Figure 12's contour illustration depicts the impacts of CR and IP. In comparison to CR, IP exerts a more significant impact on CO emissions. To attain optimal air-fuel mixing, dispersion, penetration, and swirl ratio, the fuel was effectively atomized by elevating the injection pressure. CO emissions are significantly influenced by IP and CR. Figure 13 illustrates the significant impact of IP and nanoparticle concentration on CO emissions. The incorporation of nanoparticles into B25 lowers CO emissions. An optimal method to facilitate combustion is to augment the ratio of IP to nanoparticles. In another study, increasing tomato seed methyl ester content raised BTE, and NOx but reduced CO, smoke, and HC emissions across loads. TSME–n-butanol–diesel blends offer favorable emission and performance trade-offs [26]. Figure 14 illustrates the variation in CO with respect to load and TiO_2 content. As the load increases, both combustion temperature and pressure elevate, resulting in a higher ignition flame temperature and, consequently, enhanced combustion efficiency.

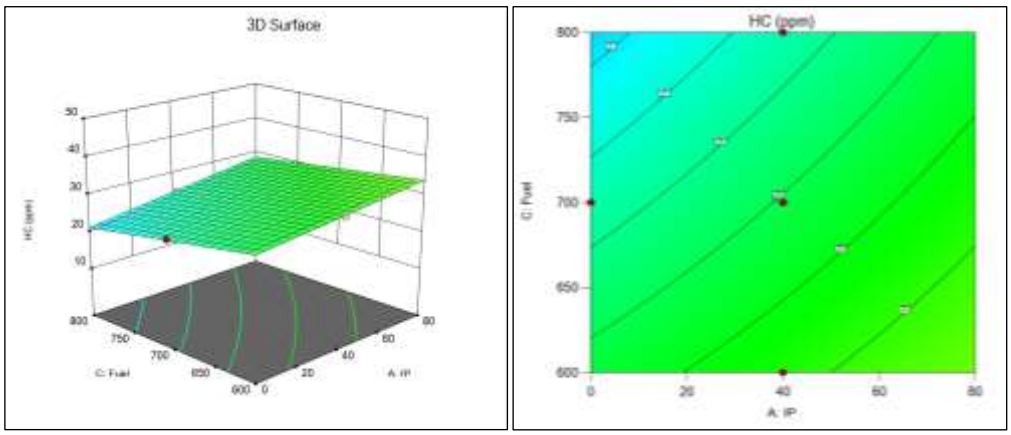


Figure 13: Three dimensional plot depicts the impact of IP and fuel and their combined impact on the CO emissions.
Source: Authors, (2026).

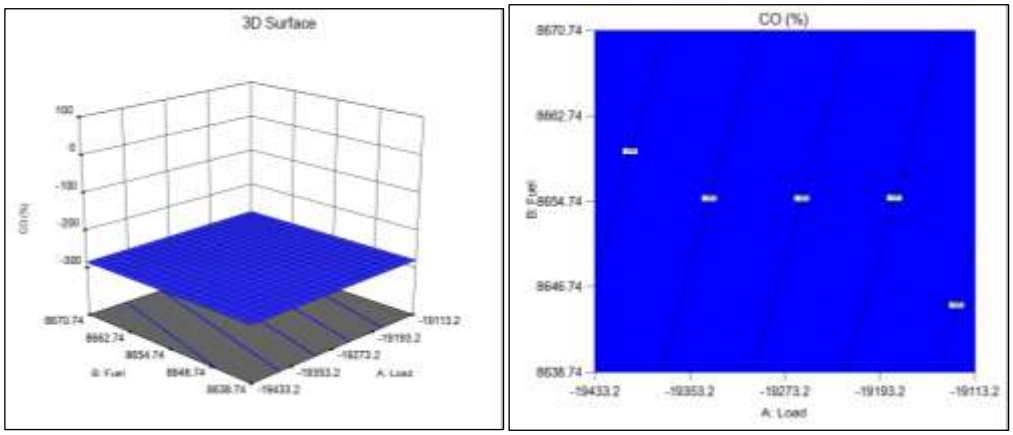


Figure 14: Three dimensional plot depicts the impact of load and fuel and their combined impact on the CO emissions.
Source: Authors, (2026).

IV.2.2 HC emissions

A second-order quadratic regression equation was created to estimate HC (utilizing the coded values), represented by equation (4).

$$HC = 28.16826\ 3.16471\ A14.+\ 73017\ B - 3.18461C0.4099\ D - 0.4195AB + 0.57291\ AC0.+\ 64735\ AD0 + .3800\ BC0.+\ 52447 + BD\text{-}0.1111\ CD \quad (4)$$

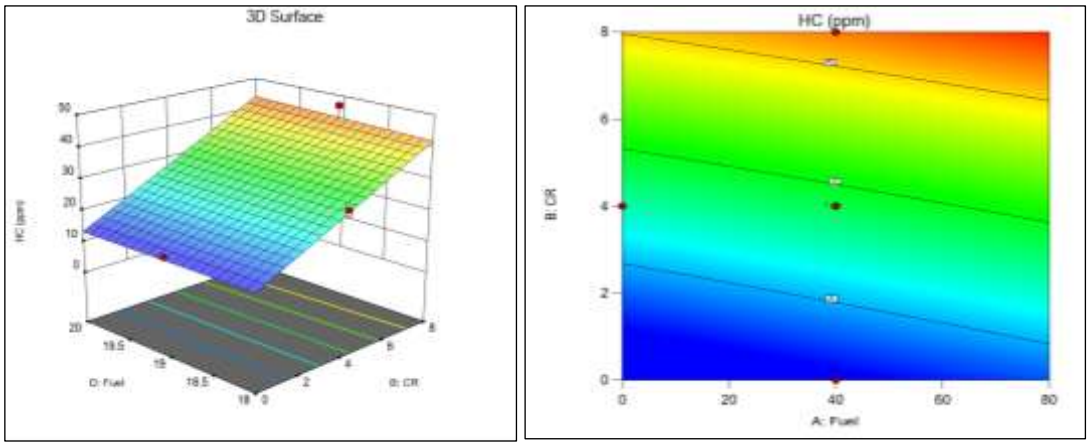


Figure 15: Three dimensional plot depicts the impact of CR and fuel and their combined impact on the HC emissions.
Source: Authors, (2026).

Elevated HC emissions result from incomplete combustion of fuel particles. At which load (4 kg) and pressure (700 bar) do the different nanoparticle concentrations in B25 and CR exhibit as HC. The outcomes are illustrated in Figure 15. Researchers found that augmenting the nanoparticle ratio resulted in a little increase in HC emissions. The large specific heat capacity of the nanoparticle is likely due to the unburned fraction of the charge being simulated. In comparison to alternative test fuels, biodiesel generated reduced hydrocarbon emissions, likely attributable to enhanced fuel ignition resulting from the elevated oxygen content in biofuel molecules. Figure 16 shows the impact of engine load and biodiesel nanoparticle ratio on hydrocarbon emissions. The presence of nanoparticles in biodiesel reduced the activation temperature of carbon and enhanced ignition, hence diminishing hydrocarbon emissions.

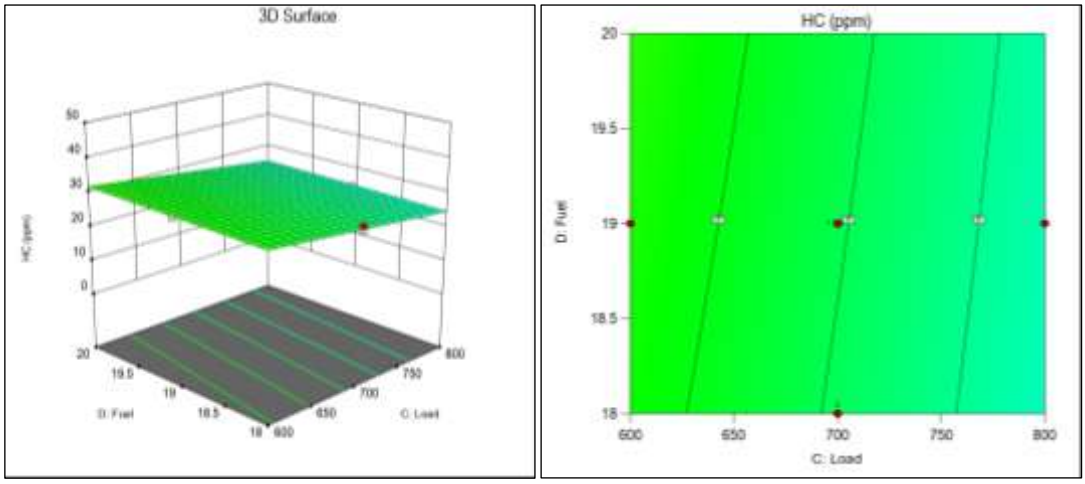


Figure 16: Three dimensional plot depicts the impact of load and fuel and their combined impact on the HC emissions.
Source: Authors, (2026).

The decrease in HC is attributed to improved fuel combustion resulting from increased IP, which facilitates optimal fuel atomization and reduces fuel droplet evaporation. Elevated compression ratios lowered HC emissions, which is advantageous for CR. An elevation in CR resulted in an increase in temperature, pressure, and flame velocity within the cylinder. Figure 17 indicated that HC emissions were reduced with increasing CR and IP. Due to the diminished flame speed during fuel ignition, flame quenching and insufficient ignition may elevate hydrocarbon levels. At elevated IP and CR, HC discharge diminished, as illustrated in Figure 17. Figure 18 illustrates the HC discharge with IP and fuel, with hold values recorded as CR (19) and load (4 kg). Consequently, an unforeseen reduction in HC levels was observed at elevated IP for the optimal nano ratio fuel ignition. A higher engine IP and biodiesel with increased nanoparticle content are effective parameters for reducing HC discharges. In another study, TiO_2 nanoparticle evaluated on Pongamia pinnata biodiesel in a single-cylinder diesel engine. TiO_2 addition improved BTE by 1.72% and BSFC by 3.57% compared to biodiesel alone, and reduced CO, HC, and NOx emissions, attributed to better atomization and oxidation from nanoparticle oxygen content [27].

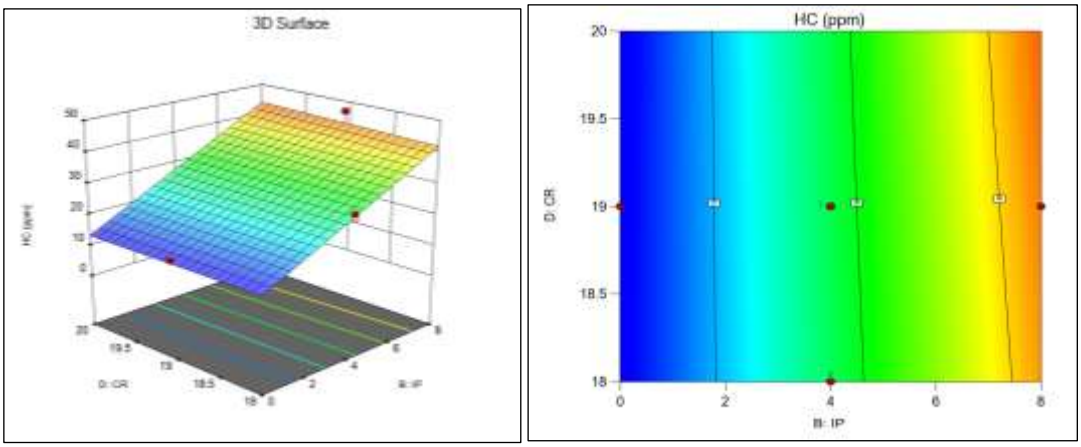


Figure 17: Three dimensional plot depicts the impact of CR and IP and their combined impact on the HC emissions.
Source: Authors, (2026).

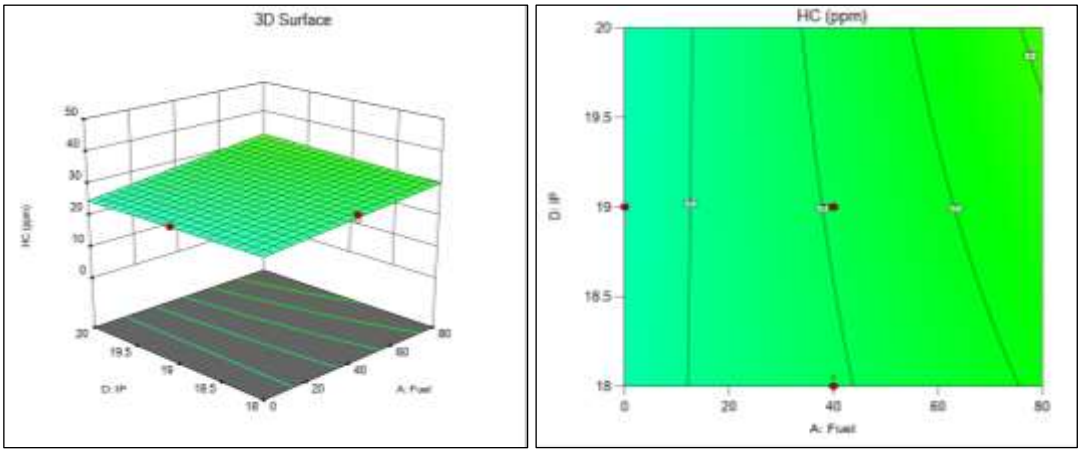


Figure 18: Three dimensional plot depicts the impact of fuel and IP and their combined impact on the HC emissions.
Source: Authors, (2026).

IV.2.3 NOx emissions

The result of the NOX estimation procedure is Equation (5), a second-order quadratic regression equation utilizing the coded values.

$$NOx = 409.8483\ 24.035\ A293 + .7823\ B112 + 4863\ C36 + .6979\ D10 + 0723\ AB8 + 6214\ AC10 + 33490\ AD0 + 43936\ BC3 + 1528\ BD - 4.4590\ CD \quad (5)$$

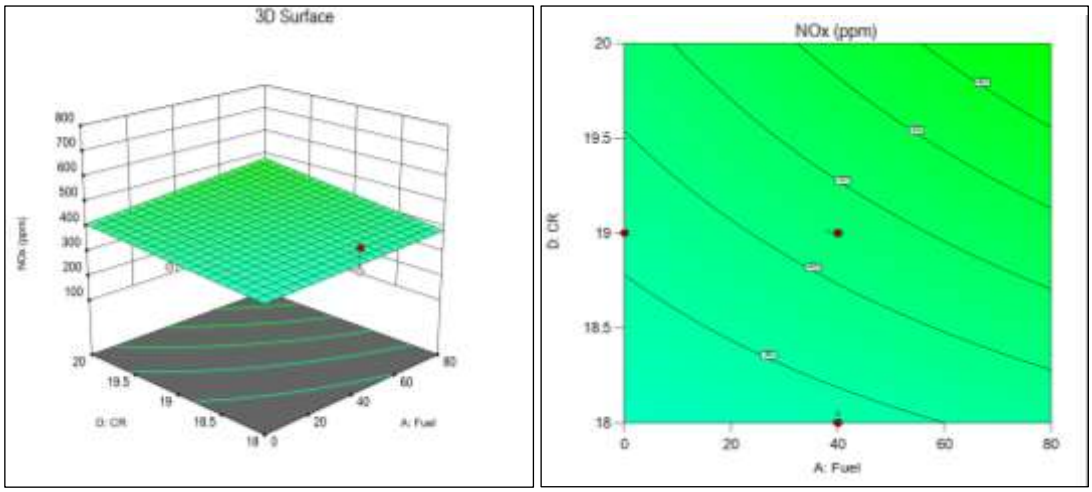


Figure 19: Three dimensional plot depicts the impact of fuel and CR and their combined impact on NOx emissions.
Source: Authors, (2026).

The combustion of biodiesel at elevated temperatures generates significant quantities of nitrogen oxides (NOx) due to its increased oxygen content. Figure 19 illustrates nitrogen oxide emissions associated with nano biofuel and CR. The rise in NOx emissions with CR may be due to elevated in-cylinder combustion pressure and temperature. Figure 20 illustrates that withhold values of load (4 kg) and injection pressure (700 bar), NOx emissions escalate at elevated compression ratios.

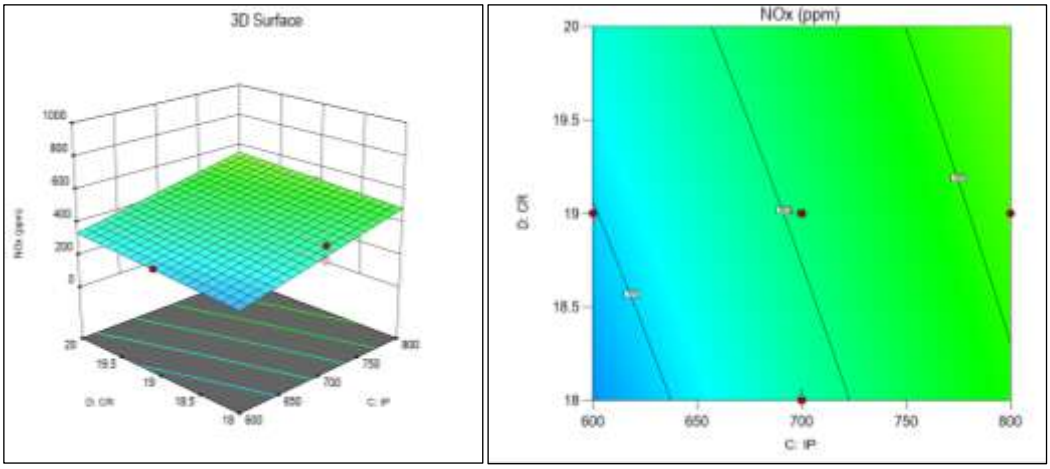


Figure 20: Three dimensional plot depicts the impact of IP and CR and their combined impact on NOx.
Source: Authors, (2026).

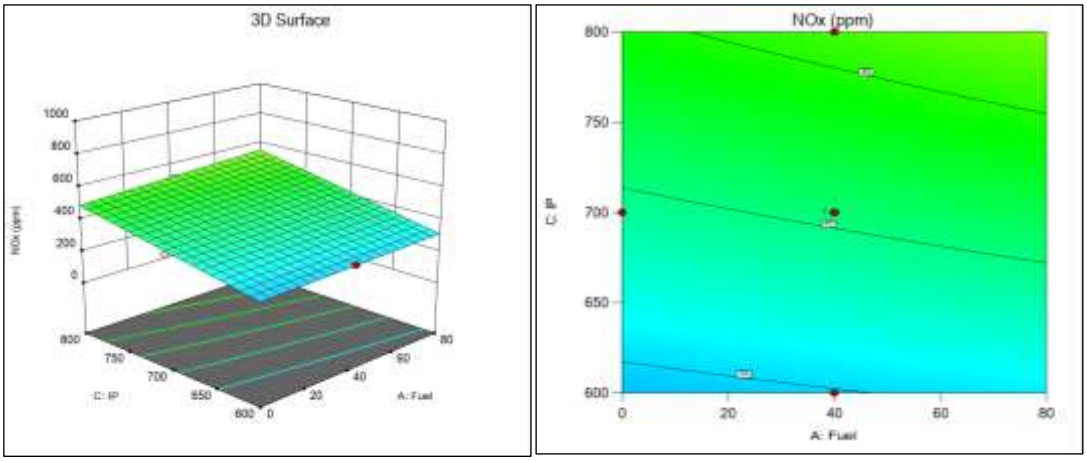


Figure 21: Three dimensional plot depicts the impact of fuel and IP and their combined impact on NOx emissions.
Source: Authors, (2026).

The contour figure indicates that increases in engine load and compression ratio corresponded with elevated NOx emissions. Figure 21 illustrates the variation of NOx emissions in relation to nano fuel and load. Figure 22 illustrates the variation of NOx emissions with varying injection pressures and nanoparticle concentrations in B25. In comparison to B25, the TiO_2 fuel blend exhibited elevated NOx emissions. Oxygenated nanoparticles improve combustion and reduce the ignition delay, since a higher ratio of nanoparticles facilitates rapid premixed ignition, elevates the burning temperature, and increases NOx emissions. In previous study, 20% *Chlorella* biodiesel blends compared with TiO_2 nanoparticle enhancements (50–150 ppm) at CRs 17–18. The D80CvBD20$TiO_2$100 blend achieved peak CP of 63.45 bar at CR 18 and reduced unburnt hydrocarbons by up to 13%. TiO_2 increased NOx but reduced CO due to higher combustion temperatures and oxygen availability [28].

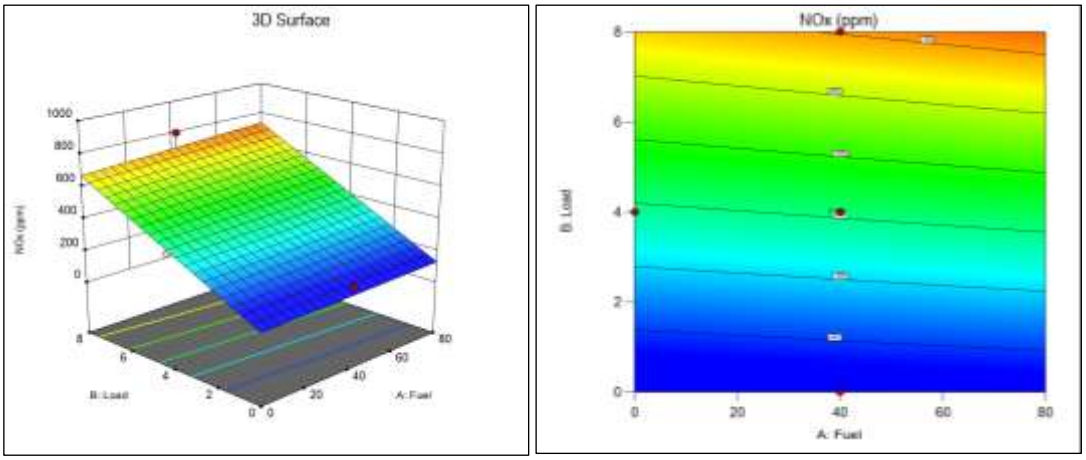


Figure 22: Three-dimensional plot depicts the impact of fuel and load and their combined impact on NOx.
Source: Authors, (2026).

Table 9: Comparative analysis of engine at optimized value.

Best value under the RSM optimization method									
Optimization method	**The best composition of the fuel**	**Engine operating parameter**			**Engine performance parameter**		**Engine emissions**		
		IP (bar)	**CR**	**Load (kg)**	**BTE (%)**	**SFC kJ/kg.hr**	**CO (%)**	**HC (ppm)**	**NO_x (ppm)**
RSM	50.47	799	20	3.63	023.96	0.45	0.044	25.03	538
The experimental result at the optimum parameter investigated by RSM									
Engine	**Composition of fuel**	**IP (bar)**	**CR**	**Load (kg)**	**BTE (%)**	**SFC kJ/kg.hr**	**CO (%)**	**HC (ppm)**	**NO_X (ppm)**
VCR CRDI engine	49.47	789	18	2.83	22.46	0.25	0.023	22.01	536
Comparative analysis between optimization result and experimental result									
Engine responses	**BTE (%)**	**SFC kJ/kg.hr**			**CO (%)**	**HC (ppm)**	**NO_x (ppm)**		
%Error	6.2	44.4			47.72	12.06	0.37		

Source: Authors, (2026).

V. CONCLUSION

This study examines the impact of various engine operating parameters on performance metrics (BTE and SFC) and emissions (CO, HC, and NOX). The parameters being examined are load, IP, CR, and the concentration of nanoparticles in biodiesel. The responses of the CRDI VCR diesel engine are optimized for each engine input parameter through response surface approach. A numerical model was developed to analyze the BTE, BSFC, CO, HC, and NOx emissions from the engine. Findings from optimization and variance analysis indicate that RSM is an effective method for refining diesel engine performance across various input parameters. The calculated R^2 values for BTE, SFC, CO, HC, and NOX emissions are 0.8990, 0.7445, 0.9016, 0.9859, and 0.9643, respectively. These figures can be accepted with confidence. The ideal input engine parameter, found by RSM, is a 50.47 nanoparticle/B25/CR/790.9 bar IP mixture operating at a 3.63 kg engine load. Moreover, for the most thoroughly examined engine inputs, the optimal values of BTE, SFC, CO, HC, and NOX are 22.46%, 0.25 kg/kWh, 23%, 22.01 ppm, and 536 ppm, respectively. Consequently, the model possesses a robust statistical base. Experimental trials indicate that errors fall within an acceptable range, and comparative analysis of RSM optimizes value. Thus, engine performance and emissions can be improved by employing the ideal value. This study's results indicate that biodiesel augmented with TiO_2 nanoparticles may serve as a cleaner and more efficient alternative fuel for engines.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Kumaran P, Sathiyaraj S
Methodology: Kumaran P and Sathiyaraj S
Investigation: Easkki Mani K and Albin Raj.
Discussion of results: Kumaran P, Sathiyaraj S and Vijayakumar K.
Writing – Original Draft: Kumaran P
Writing – Review and Editing: Sathiyaraj S and Vijayakumar K.
Resources: Kumaran P
Supervision: Sathiyaraj S and Vijayakumar K.
Approval of the final text: Kumaran P, Sathiyaraj S and Vijayakumar K.

VII. REFERENCES

[1] F. Markić, K. Kraljić, V. Stulić, S. Pleslić, T. V Pavičić, and N. Maltar-Strmečki, "Improving the extraction of tomato seed oil and the retention of bioactive substances using pulsed electric field technology," Future Foods, vol. 12, 2025, doi: 10.1016/j.fufo.2025.100706.

[2] P. Squillace, F. Adani, and B. Scaglia, "Supercritical CO2 extraction of tomato pomace: Evaluation of the solubility of lycopene in tomato oil as limiting factor of the process performance," Food Chem, vol. 315, 2020, doi: 10.1016/j.foodchem.2020.126224.

[3] K. Sangeetha, R. B. Ramyaa, A. Mousavi Khaneghah, and M. Radhakrishnan, "Extraction, characterization, and application of tomato seed oil in the food industry: An updated review," J Agric Food Res, vol. 11, 2023, doi: 10.1016/j.jafr.2023.100529.

[4] I. Solaberrieta, A. C. Mellinas, J. Espagnol, M. Hamzaoui, A. Jiménez, and M. C. Garrigós, "Valorization of Tomato Seed By-Products as a Source of Fatty Acids and Bioactive Compounds by Using Advanced Extraction Techniques," Foods, vol. 11, no. 16, 2022, doi: 10.3390/foods11162408.

[5] R. Karami, M. G. Rasul, and M. M. K. Khan, "An Empirical and Computational Fluid Dynamics Analysis of Combustion Performance of a Diesel Engine Fueled With Tomato Seed Oil Biodiesel," Journal of Energy Resources Technology, Transactions of the ASME, vol. 145, no. 4, 2023, doi: 10.1115/1.4055470.

[6] D. P. Rao et al., "Adsorptive removal of cadmium from electroplating wastewater using hybrid composite of thiol-grafted seed gum of Tamarindus indica and Teff hay biocarbon," Zeitschrift fur Physikalische Chemie, vol. 239, no. 5, pp. 633–654, 2025, doi: 10.1515/zpch-2024-0715.

[7] B. Gaddala et al., "Evaluation of detoxification performance of Pithecellobium dulce seed pod powder upon acid treatment: removal of reactive red 195-A dye and kinetic analysis," Zeitschrift fur Physikalische Chemie, vol. 239, no. 6, pp. 885–910, 2025, doi: 10.1515/zpch-2024-0879.

[8] T. Ouatmani, H. Haddadi-Guemghar, L. Boulekbache-Makhlouf, D. Mehidi-Terki, A. Maouche, and K. Madani, "A sustainable valorization of industrial tomato seeds (cv Rio Grande): Sequential recovery of a valuable oil and optimized extraction of antioxidants by microwaves," J Food Process Preserv, vol. 46, no. 1, 2022, doi:

10.1111/jfpp.16123.

[9] P. Kumaran, S. Natarajan, R. Shanmuga Raj, S. Dhanaraj, and V. R. Kumar, “Performance and emissions on VCR diesel engine with turbocharger setup running using tomato seed oil,” in Materials Today: Proceedings, 2020, pp. 6078–6082. doi: 10.1016/j.matpr.2020.10.080.

[10] P. Kumaran, S. Natarajan, M. Maheshwaran, K. Rizal, and B. C. Rohith, “Investigation on Performance and Emission Characteristics of VCR Diesel Engine with Different Nozzles Using Blends of Tomato Seed Oil,” in AIP Conference Proceedings, 2023. doi: 10.1063/5.0110969.

[11] H. A. Mahmood, A. O. Al-Sulttani, H. A. Alrazen, and O. H. Attia, “The impact of different compression ratios on emissions, and combustion characteristics of a biodiesel engine,” AIMS Energy, vol. 12, no. 5, pp. 924–945, 2024, doi: 10.3934/ENERGY.2024043.

[12] M. Kiralan and O. Ketenoglu, “Utilization of Tomato (Solanum lycopersicum) by-Products: An Overview,” in Mediterranean Fruits Bio-wastes: Chemistry, Functionality and Technological Applications, 2022, pp. 799–818. doi: 10.1007/978-3-030-84436-3_34.

[13] M. Casa, C. Prizio, and M. Miccio, “Biodiesel production from tomato seed by transesterification with Alkaline and ‘Green’ catalysts: Simulation and discussion,” Chem Eng Trans, vol. 87, pp. 451–456, 2021, doi: 10.3303/CET2187076.

[14] R. Karami, M. G. Rasul, M. M. K. Khan, and M. Mahdi Salahi, “Experimental and computational analysis of combustion characteristics of a diesel engine fueled with diesel-tomato seed oil biodiesel blends,” Fuel, vol. 285, 2021, doi: 10.1016/j.fuel.2020.119243.

[15] V. H. Özyurt, I. Tetik, and S. Ötleş, “Influence of process conditions on ultrasound-assisted protein extraction from cold pressed tomato seed waste,” J Food Process Preserv, vol. 45, no. 12, 2021, doi: 10.1111/jfpp.16079.

[16] V. H. Ozyurt, B. Çakaloğlu, and S. Otles, “Optimization of cold press and enzymatic-assisted aqueous oil extraction from tomato seed by response surface methodology: Effect on quality characteristics,” J Food Process Preserv, vol. 45, no. 5, 2021, doi: 10.1111/jfpp.15471.

[17] B. Maleki, H. Esmaeili, M. Mansouri, D. Kumar, and B. Singh, “Enhanced conversion of dairy waste oil to biodiesel via novel and highly reactive UiO-66-NH2/ZnO/TiO2 nano-catalyst: Optimization, kinetic, thermodynamic and diesel engine studies,” Fuel, vol. 339, 2023, doi: 10.1016/j.fuel.2022.126901.

[18] S. Muniyappan and R. Krishnaiah, “Experimental and artificial neural network (ANN) study on the impact of three different metal oxide nanoparticle combustion enhancers enriched on mahua biodiesel-diesel fuelled CI engine,” Engineering Research Express, vol. 6, no. 4, 2024, doi: 10.1088/2631-8695/ad9984.

[19] S. Muniyappan and R. Krishnaiah, “The influence of TiO2 nanoparticles on the performance, combustion, and emissions on ternary blends of n-heptane, mahua biodiesel and diesel-fuelled engine using response surface methodology,” Case Studies in Chemical and Environmental Engineering, vol. 10, 2024, doi: 10.1016/j.cscee.2024.100900.

[20] X. Zhang, “RSM-based optimization for ultrasound-assisted transesterification of Jojoba oil to biodiesel via ZnFe2O4@TiO2 retrievable nanocatalyst: Application on the diesel engine,” Ind Crops Prod, vol. 220, 2024, doi: 10.1016/j.indcrop.2024.119144.

[21] P. Sujin, S. S. Selva Roji, A. J. Kings, and L. R. M. Miriam, “Optimization of engine performance, emission and combustion parameters by using mixed nonedible oil biodiesel with nano-additives using hybrid techniques,” Energy, vol. 305, 2024, doi: 10.1016/j.energy.2024.132413.

[22] A. P. Edlabadkar et al., “Optimization of Mechanical Properties of Cordia dichotoma/Alumina Nanoparticle Reinforced Epoxy Nanocomposites Under Alkali Treatment Using ANN Techniques,” Journal of Environmental Nanotechnology, vol. 14, no. 2, pp. 525–540, 2025, doi: 10.13074/jent.2025.06.2521549.

[23] R. Karami, M. G. Rasul, and M. M. K. Khan, “CFD simulation and a pragmatic analysis of performance and emissions of tomato seed biodiesel blends in a 4-cylinder diesel engine,” Energies (Basel), vol. 13, no. 14, 2020, doi: 10.3390/en13143688.

[24] K. Lenin, S. Raghuvaran, K. B. Hariharan, M. Manimaran, H. Ramakrishnan, and S. Dhanabalan, “Experimental investigation on single cylinder four-stroke diesel engine with copper oxide nano-additives using tomato seed oil,” in Materials Today: Proceedings, 2022, pp. 1280–1291. doi: 10.1016/j.matpr.2022.08.401.

[25] M. Sunil Kumar, R. Rajasekar, S. Ganesan, S. P. Venkatesan, and V. Praveen Kumar, “Evaluation of metal oxide nano particles in lemongrass biodiesel for engine performance, emission and combustion characteristics,” in Materials Today: Proceedings, 2021, pp. 3657–3665. doi: 10.1016/j.matpr.2020.10.796.

[26] H. Sivasubramanian, V. Sundaresan, S. K. Ramasubramaniam, S. R. Shanmugaiah, and S. K. Nagarajan, “Investigation of biodiesel obtained from tomato seed as a potential fuel alternative in a CI engine,” Biofuels, vol. 11, no. 1, pp. 57–65, 2020, doi: 10.1080/17597269.2017.1338124.

[27] M. K. Parida, P. Mohapatra, S. S. Patro, and S. Dash, “Effect of TiO2 nano-additive on performance and emission characteristics of direct injection compression ignition engine fueled with Karanja biodiesel blend,” Energy Sources, Part A: Recovery, Utilization and Environmental Effects, vol. 46, no. 1, pp. 7521–7530, 2024, doi: 10.1080/15567036.2020.1756991.

[28] V. Hariram et al., “Influence of TiO2 Nanoparticles on Engine Parameters at Varied Compression Ratio of CI Engine Fuelled with Biodiesel,” Journal of Environmental Nanotechnology, vol. 14, no. 2, pp. 58–68, 2025, doi: 10.13074/jent.2025.06.2511378.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 259-264. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2881

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

IMPLEMENTATION AND COMPARATIVE ANALYSIS OF A FUZZY LOGIC CONTROLLER FOR INDUCTION MOTOR VECTOR CONTROL

Salim Mahdab*[1], Mohamed Fekir[2]

[1]Department of Electronic and telecommunication, University of Djilali Bounaama, Khemis Miliana, Ain Defla, Algeria.
[2]Department of Electrical Engineering, University of Djilali Bounaama, Khemis Miliana, Ain Defla, Algeria

[1]https://orcid.org/0009-0007-8691-977X, [2]http://orcid.org/0000-0001-9743-9543

Email: * s.mahdab@univ-dbkm.dz, m.fekir@univ-dbkm.dz

ARTICLE INFO

Article History
Received: November 1, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Indirect Vector Control (IVC),
PI Controller,
Fuzzy Logic Controller (FLC)

ABSTRACT

Indirect vector control of induction motor (IM) drives operates on the principle of decoupling the stator current into independent torque and flux-producing components. This paper presents the implementation of a Fuzzy Logic Control (FLC) scheme for an IM, based on its dynamic d-q model. An intelligent Fuzzy Logic Controller is designed using a knowledge-based rule set to achieve robust and efficient drive control. The performance of the proposed FLC is rigorously compared against a conventional fixed-gain Proportional-Integral (PI) controller. Key performance metrics include settling time and the dynamic response to sudden load disturbances. Furthermore, the harmonic spectrum of the output current is evaluated for both control strategies. The steady-state performance of the IM drive is analyzed, and comparative results are presented to demonstrate the superiority of the fuzzy logic-based approach.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

For over a century, induction motors (IMs) have been a cornerstone of industrial electromechanical conversion due to their exceptional simplicity, rugged construction, and high efficiency. The invention of this asynchronous machine significantly propelled the transition from direct current (DC) to alternating current (AC) systems in the generation, transmission, and distribution of electrical energy. A primary advantage of the induction motor is the absence of sliding electrical contacts, which results in a robust and maintenance-friendly design[1]. Despite decades of research into efficient speed control methodologies for induction motors, solutions available until recent years remained unsatisfactory, often compromising on complexity, cost, efficiency, or dynamic performance.

The advancement of semiconductor technology over the past two decades has provided the necessary impetus, making static frequency converters economically viable and enabling widespread adoption of adjustable-speed drives for numerous applications. Control complexity presents a fundamental challenge. Separately excited DC drives benefit from independent control of flux and torque, leading to a relatively simple control structure [2]. In contrast, induction motors require the coordinated, nonlinear control of both the magnitude and phase of the stator current. This complexity can be mitigated through vector control, or field-oriented control (FOC). This technique resolves the stator flux linkages onto a rotating reference frame, effectively decoupling the flux and torque-producing components.

By requiring the instantaneous position of the flux linkages, FOC enables control of the AC induction motor that is analogous to the control of a separately excited DC motor. Conventional drive systems typically employ fixed-gain Proportional-Integral (PI) controllers .However, these controllers exhibit significant limitations, including sensitivity to parameter variations and poor disturbance rejection, often resulting in prolonged settling times following load transients. Consequently, numerous adaptive control techniques have been proposed to enhance robustness and performance[3]. In recent years, fuzzy logic control (FLC) has emerged as a powerful methodology for managing complex, nonlinear systems. Fuzzy set theory has fostered active and fruitful research, leading to its successful application in motor control.

Prior works present comprehensive surveys of FLC and describe general methodologies for its construction and performance assessment. Furthermore, discusses comprehensive procedures and guidelines for defining input parameters and constructing fuzzy logic rule bases. Additional methodologies for the design and theoretical analysis of FLCs are proposed in including explanations of the relationship between control resolution and the fuzzification of variables, as well as guidance for designing and tuning scaling gains[4]. In this context, the present paper proposes the design and implementation of an intelligent controller based on Fuzzy Logic for induction motor drives. The principal advantage of an FLC is its independence from a precise mathematical model of the plant; instead, it leverages a foundation of IF-THEN linguistic rules derived from expert knowledge and operational experience[5].

II. INDUCTION MOTOR MODELING

The dynamic analysis and modeling of rotating field machines are underpinned by well-established theoretical frameworks. This work considers a standard induction motor model characterized by a uniform air gap and a sinusoidally distributed magnetomotive force (mmf). For analytical simplicity, the effects of magnetic saturation and variations in machine parameters are neglected.The dynamic model of the induction motor is developed using a reference frame transformation. The three-phase electrical quantities (abc-frame) are transformed into an equivalent two-phase system within a direct and quadrature (d-q) rotating reference frame. This transformation is derived based on the principle of power invariance, which necessitates that the total power calculated in the two-phase model must be identical to that of the original three-phase system.The magnetomotive forces (mmfs) along the d and q axes are obtained by mathematically resolving the mmfs of the three stator phases onto these orthogonal axes. The formal mathematical relationship governing the transformation between the d-q axes currents and the a-b-c phase currents is given as follows:

$$i_{ds} = \frac{2}{3} V_d \left(i_{as} - \frac{i_{bs}}{2} - \frac{i_{cs}}{2}\right) \tag{1}$$

$$i_{qs} = \frac{1}{\sqrt{3}} \left(i_{cs} - i_{bs}\right) \tag{2}$$

Where ias, ibs and ics are the three phase currents and Vd is the dc link voltage at the inverter input. These equations are also applicable to the voltage and flux linkage transformation Applying this transformation to the three phase quantities . The electrical model of the induction motor is expressed by the matrix equation (3) in stationary reference frame.

$$\begin{pmatrix} v_{ds} \\ v_{qs} \\ v_{dr} \\ v_{qr} \end{pmatrix} = \begin{pmatrix} R_s + L_s p & 0 & L_m p & 0 \\ 0 & R_s + L_s p & 0 & L_m p \\ L_m p & \omega_r L_m & R_r + L_r p & \omega_r L_r \\ -\omega_r L_m & L_m p & -\omega_r L_r & R_r + L_r p \end{pmatrix} \begin{pmatrix} i_{ds} \\ i_{qs} \\ i_{dr} \\ i_{qr} \end{pmatrix} \tag{3}$$

Where vds, vqs, ids, iqs Rs, Ls, Rr, Lr, Lm and θr are the d-q axes voltages and currents, stator resistance, stator inductance, rotor resistance, rotor inductance, mutual inductance betweenthe stator and rotor windings and the rotor position respectively. The stator and rotor flux linkages in the synchronous reference frames are defined as:

$$\Phi_s = L_s \bar{I}_s + L_m \bar{I}_r$$
$$\Phi_r = L_r \bar{I}_r + L_m \bar{I}_s \tag{4}$$

Where Ψ_s, Ψ_r, I_s and I_r are the stator and rotor flux-linkage Space vector and the stator and rotor current space vector, all in stationary reference frame. The electromagnetic torque obtained from machine flux linkages and currents is as:

$$T_e = \frac{3}{2}\frac{P}{2} L_m \left(i_{qs} \Phi_{dr} - i_{ds} \Phi_{qr}\right) \tag{5}$$

Where Te, P, Φdr, Φqr are the electromagnetic torque, number of poles, rotor d-q axes fluxes respectively. The Electromagnetic dynamic equation describing the mechanical model of the induction motor is given by:

$$T_e = J \frac{\omega_n}{dt} + T_l + B\omega_n \tag{6}$$

Where J, T1, B, ωm are the moment of inertia of motor and Load , the load torque, the friction coefficient and the mechanical speed. The equations (3) and (6) form the mathematical model equations of a three phase induction motor.

II.1 INDIRECT VECTOR CONTROL

The structure of the indirect vector control method is illustrated in Figure 1. In contrast to the direct vector control approach, this method operates as a feedforward control scheme, as it relies on the indirect estimation of the slip speed for field orientation

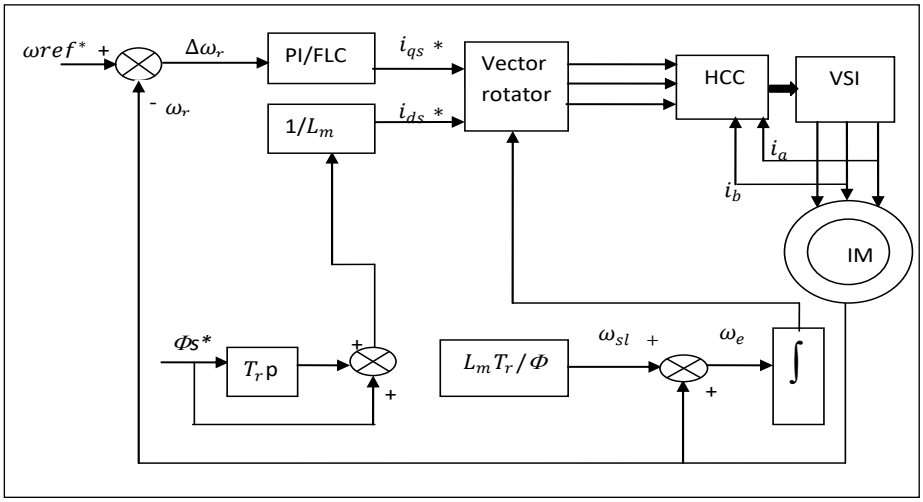


Figure 1: Indirect Vector Control Scheme.
Source: Authors, (2026).

The speed error, with the help of a PI controller or any other intelligent controller, is converted into a torque controlling current component iqs*, of the stator current. This current component t is used to regulate the torque along with the slip speed

$$i^*_{qs} = k_p \Delta\omega_r + k_i \int \Delta\omega_r dt \tag{7}$$

Similarly the flux producing current component ids*, is obtained from the stator flux linkage reference value and is given by the following equation:

$$\Phi^*_s = \frac{R_r L_m}{R_r + Lp} i^*_{ds} \tag{8}$$

$$T_e = \frac{3}{2}\frac{P}{2}\frac{L_m}{L_r}\Phi^*_s i^*_{qs} \tag{9}$$

$$\omega_{sl} = \omega_e - \omega_r = \frac{R_r}{L_r}\frac{i^*_{qs}}{i^*_{ds}} \tag{10}$$

The slip speed, together with the feedback rotor speed, is integrated to obtain the stator reference flux linkage space vector position θe.

$$O_e = \int (\omega_{sl} + \omega_r) dt \tag{11}$$

The position of the stator flux space vector is utilized to transform the two-phase direct and quadrature (d-q) axis current components back into three-phase quantities. These transformed currents serve as the reference command signals. The actual three-phase currents measured from the induction motor are continuously compared against these reference values. The resulting current errors are processed by two-level hysteresis band controllers. The hysteresis controllers regulate the motor currents by constraining them within a predefined hysteresis band. This control strategy ensures the machine achieves its required dynamic performance. Furthermore, the flux reference is maintained constant at its rated value for operation up to the base speed of the motor.

III. FUZZY LOGIC CONTROLLER

Fuzzy Logic implementation requires no exact knowledge of A model. The block diagram implementation of a FLC is shown in Figure 2. It involves the use of the concept of fuzzy subset and rule based modeling. By permitting certain amount of imprecision, complex solutions are modeled with ease[6].

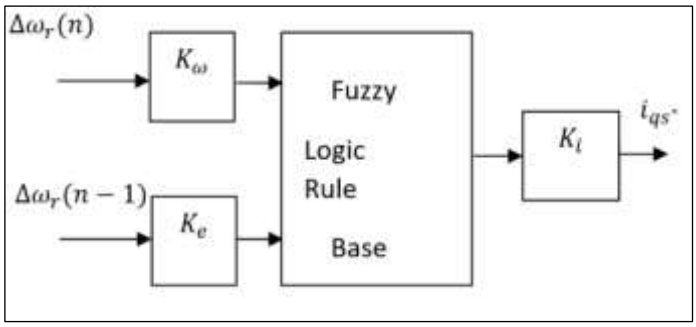


Figure 2: Fuzzy Logic Based Controller block.
Source: Authors, (2026).

III. 1 CONCEPT OF FUZZY LOGIC

The idea of formulating the control algorithms by logical rules introduced the implementation of human understanding and human thinking in control algorithms. The lack of analytical description makes the fuzzy control conceptually different from conventional control [7-9]. One defines a subset with the aid of a characteristic function which describes membership in the subset. Let X be the universe of discourse and let S b e a subset of X. The characteristic function associated with S is a mapping

$$f_s: X \rightarrow \{0,1\} \tag{12}$$

In the framework of fuzzy set theory the characteristic function is called the membership function associated with the fuzzy subset A. A fuzzy logic controller is describe d by a knowledge-based system consisting of IF. THEN rules with vague predicates and a fuzzy logic inference mechanism. The present fuzzy logic controller adopted is of the Mamdani Controller type[10].The speed error and the change in the speed error are given as inputs to the FLC. The torque producing current component t is the output. A knowledge rule base which simulates the performance of the system is defined. The rule base acts upon the inputs to produce the given outputs.

III. 2 CONCEPT MEMBERSHIP FUNCTION

The Fuzzy Logic Controller initially converts the crisp error and change in error variables into fuzzy variables and then are mapped into linguistic labels. Membership functions are associated with each label consists of two inputs and one output. The linguistic labels are divided into seven groups. They are:NH-negative high, N-Negative, NL-negative low, Z-zero, PL-positive low, P-positive, PH-positive high. Each of the Inputs and the output contain membership functions with all these seven linguistics. This method of formulation of control algorithms allows implementing heuristic strategies [11-14]. A straightforward source of deriving the linguistic control strategies are human experience and understanding, which essentially contain the model of the control system in an implicit form.

III. 3 CONCEPT KNOWLEDGE RULE BASE

The mapping of the fuzzy inputs into the required output is derived with the help of a rule base as given in Table 1. Each rule of the FLC is characterized with an IF part, called the antecedent, and with a THEN part called the consequent. The antecedent of a rule contains a set of conditions and the consequent contains a conclusion. If the conditions of the antecedents are satisfied, then the conclusions of the consequent apply. Considering the first rule, it can be interpreted as: IF change in speed error is NH and change is speed is NH, THEN the output will be NH.

Table 1: Knowledge Rule Base.

E / CE	NH	N	NL	Z	PL	P	PH
NH	NH	NH	NH	NH	NL	NL	Z
N	NH	NH	N	N	N	Z	PL
NL	NH	N	NL	NL	Z	PL	P
Z	NH	N	NL	Z	PL	P	PH
PL	N	NL	Z	PL	PL	P	PH
P	NL	Z	PL	P	P	PH	PH
PH	Z	PL	P	PH	PH	PH	PH

Source: Authors, (2026).

III. 4 DEFUZZIFICATION

Generally the output obtained is fuzzy in nature and has to be Converted into a crisp value by using any Defuzzification technique.

IV. RESULTS AND DISCUSSIONS

In the following, we will present the results of the essential parameters of our machine, we will be particularly interested in the currents of the machine as well as the torque and the speed to highlight the efficiency of our fuzzy controller compared to a classic PI controller. The current response of the machine for both case i.e. PI and FLC is presented in the Fig.3 and Fig.4

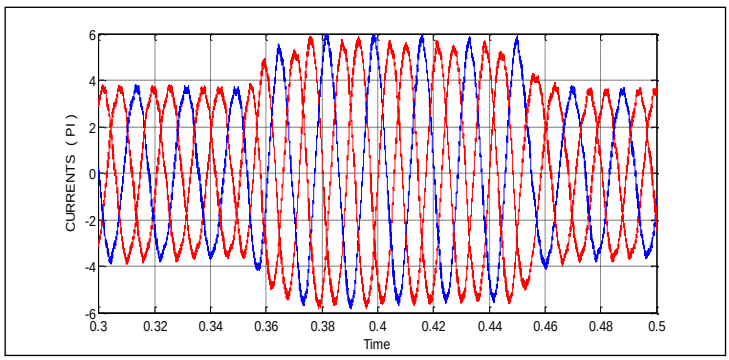


Figure 3: Stator currents response for PI.
Source: Authors, (2026).

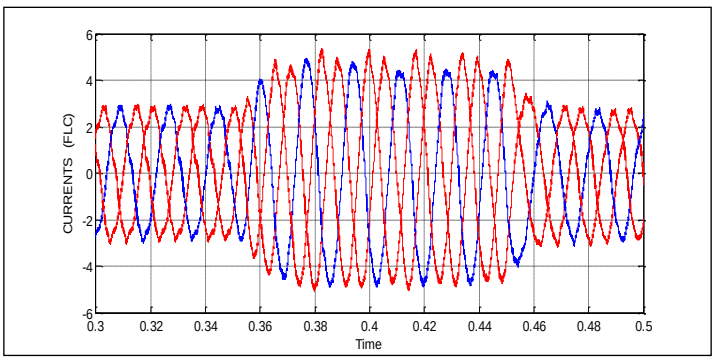


Figure 4: Stator currents response for FLC.
Source: Authors, (2026).

Figure 3, 4 presents a comparative analysis of the machine's current response under both Proportional-Integral (PI) and Fuzzy Logic Control (FLC) schemes. The graphical results clearly indicate that the FLC strategy exhibits significantly greater robustness to load disturbances compared to the conventional PI controller. The torque response for both cases of classical controllers is shown simultaneously in Figure 5.

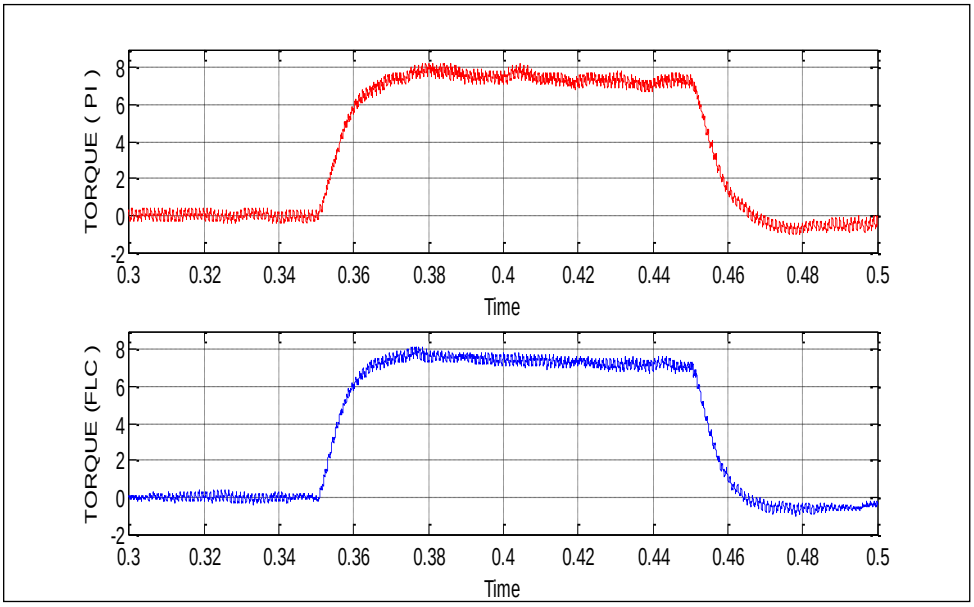


Figure 5: Torque response for PI and FLC.
Source: Authors, (2026).

The waveform indicates that the Fuzzy Logic Controller (FLC) achieves a significant reduction in current ripple and a markedly quicker dynamic response to load transients compared to the PI controller. Figure 6 (a,b) shows the speed response for both PI and fuzzy controllers.

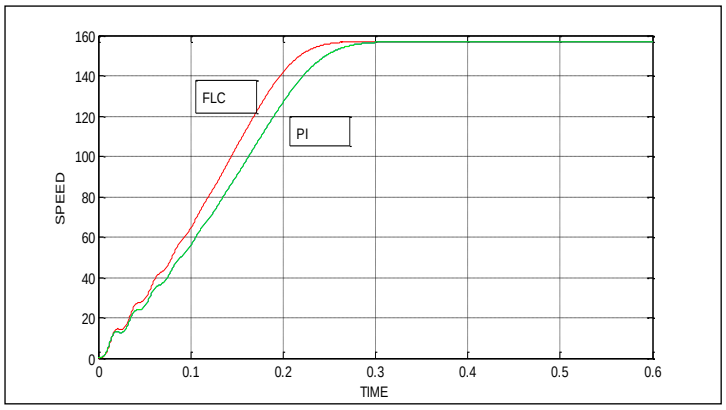


Figure 6 (a): Speed response for PI and FLC.
Source: Authors, (2026).

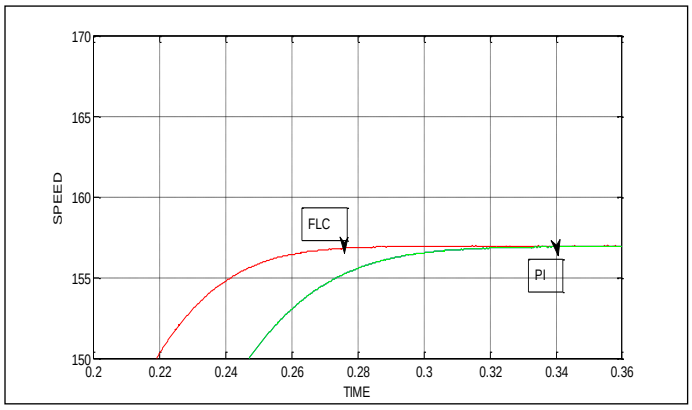


Figure 6 (b): Speed response for FLC (0.275sec) and PI(0.34sec).
Source: Authors, (2026).

When a load is applied at 0.35 sec the speed response for Both controller are as shown in Fig.6 (c),

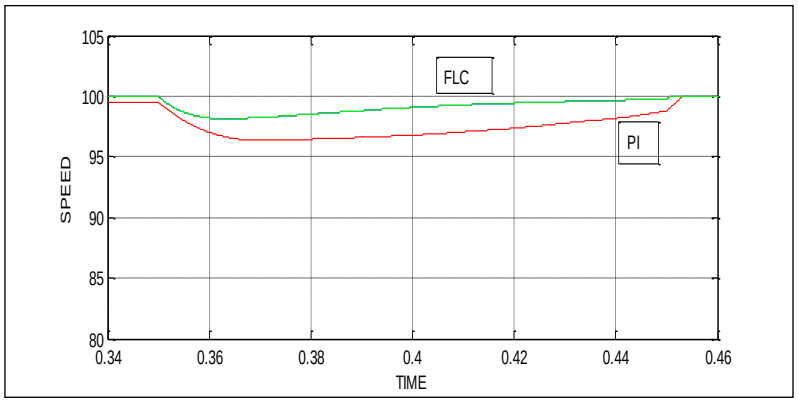


Figure 6 (c): Speed response for FLC and PI with andWithout load impact .
Source: Authors, (2026).

The dynamic response to a load torque application at t = 0.35 s is detailed in Fig. 6(c). This analysis confirms that the FLC strategy results in a smaller speed deviation and a quicker recovery compared to the PI controller.

V. CONCLUSIONS

This paper presents the implementation of a Fuzzy Logic Controller (FLC) for an indirect vector-controlled induction motor drive. The drive system was simulated using both a conventional Proportional-Integral (PI) controller and the proposed FLC to facilitate a comparative analysis. The performance of both controllers was evaluated and compared under dynamic operating conditions. Simulation results demonstrate that the FLC-based strategy achieves a significantly faster settling time and superior overall performance compared to the PI controller.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Salim Mahdab and Mohamed Fekir.
Methodology: Salim Mahdab.
Investigation: Salim Mahdab and Mohamed Fekir.
Discussion of results: Salim Mahdab.
Writing – Original Draft: Salim Mahdab.
Writing – Review and Editing: Salim Mahdab and Mohamed Fekir.
Resources: Salim Mahdab.
Supervision: Author Salim Mahdab.
Approval of the final text: Author Salim Mahdab and Mohamed Fekir.

VII. REFERENCES

[1] S. M. A. Mohammadi, M. R. Alizadeh Pahlavani et A. M. D. A. L. A. S. I. A. Vahedi, "A Comparative Study of PI and Fuzzy Logic Controllers in Field-Oriented Control of Induction Motor Drives for Electric Vehicle Application," IEEE Transactions on Vehicular Technology, vol. 71, no. 6, pp. 5878-5892, June 2022.

[2] B. K. Bose, *Modern Power Electronics and AC Drives*. Upper Saddle River, NJ, USA: Prentice Hall, 2002.

[3] W. Leonhard, *Control of Electrical Drives*, 3rd ed. Berlin, Germany: Springer-Verlag, 2001.

[4] P. Vas, *Sensorless Vector and Direct Torque Control*. London, U.K.: Oxford University Press, 1998.

[5] I. Takahashi and T. Noguchi, "A new quick-response and high-efficiency control strategy of an induction motor," *IEEE Trans. Ind. Appl.*, vol. IA-22, no. 5, pp. 820-827, Sep. 1986.

[6] C. T. Lin, C. S. G. Lee, and C. C. Lee, "Neural fuzzy systems: A neural-fuzzy synergism to intelligent systems," *Prentice-Hall, Upper Saddle River, NJ*, 1996.

[7] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338-353, Jun. 1965.

[8] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *Int. J. Man-Mach. Stud.*, vol. 7, no. 1, pp. 1-13, Jan. 1975.

[9] R. Krishan and S. Singh, "FPGA based fuzzy logic controller for electric vehicle," in *Proc. IEEE Int. Conf. Power Electron., Drives and Energy Systems (PEDES)*, 2014, pp. 1-6.

[10] M. N. Uddin, T. S. Radwan, and M. A. Rahman, "Performances of fuzzy-logic-based indirect vector control for induction motor drive," *IEEE Trans. Ind. Appl.*, vol. 38, no. 5, pp. 1219-1225, Sep./Oct. 2002.

[11] T. Orlowska-Kowalska and M. Dybkowski, "Stator-current-based MRAS estimator for a wide range speed-sensorless induction-motor drive," *IEEE Trans. Ind. Electron.*, vol. 57, no. 4, pp. 1296-1308, Apr. 2010.

[12] J. O. P. Pinto, B. K. Bose, L. E. B. da Silva, and M. P. Kazmierkowski, "A neural-network-based space-vector PWM controller for voltage-fed inverter induction motor drive," *IEEE Trans. Ind. Appl.*, vol. 36, no. 6, pp. 1628-1636, Nov./Dec. 2000.

[13] MathWorks, "Fuzzy Logic Toolbox™ User's Guide," Natick, MA, USA, 2022. [En ligne]. Disponible: https://www.mathworks.com/help/fuzzy/

[14] Texas Instruments, "InstaSPIN™ Motor Control Solutions," Dallas, TX, USA. [En ligne].

APPENDIX

The parameters of 3 phase Induction Motor, in SI units are:

P = 1.5kW,
Rs = 4.85Ω,
Rr = 3.805Ω
Ls = 0.274H,
Lr = 0.274H,
Lm = 0.258H
J = 0.031kg.m2,
f = 0.001136Nm/rd/s

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 265-276. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2884

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

ENHANCED ECONOMIC DISPATCH OPTIMIZATION USING MOMENTUM IN THE ORCA PREDATION ALGORITHM

Vivi Aida Fitria*[1], Arif Nur Afandi [2], Apriharta[3] , Azwar Riza Habibi[4]

[1,4] Informatics, Faculty of Technology and Design, Institut Teknologi dan Bisnis Asia, Malang, Indonesia
[2, 3] Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, Indonesia.

[1]https://orcid.org/0009-0006-2941-6888, [2]https://orcid.org/0000-0001-9019-810X
[3]https://orcid.org/0000-0002-5313-6978, [4]https://orcid.org/0009-0003-8200-7635

Email: *viviaida@asia.ac.id, vivi.aida.2205349@students.um.ac.id

ARTICLE INFO

Article History
Received: November 3, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Self-Adaptive Momentum Algorithm,
Economic Dispatch Optimization,
Metaheuristic Algorithms,
Orca Predation Algorithm,
Power System Optimization.

ABSTRACT

The economic dispatch problem aims to minimize total generation costs while maintaining a balanced power allocation in the electrical power system. This paper introduces a new algorithm called the Self-Adaptive Momentum Orca Predation Algorithm (SAMOPA), the first variant of OPA that integrates an adaptive momentum mechanism to solve the economic dispatch problem. This integration enhances the balance between exploration and exploitation, accelerates the convergence process, and reduces the risk of getting stuck in local solutions. SAMOPA was tested on three economic dispatch scenarios with 6, 15, and 40 power plants. In the 6-unit system, SAMOPA achieved the best cost of 15275.9304 dollars with a standard deviation of 2.54E-06, lower than OPA's deviation of 0.0029. SAMOPA's computation time was also faster, at 0.27 seconds compared to OPA's 0.9751 seconds. For the 15-unit system, SAMOPA reduced total generation costs by 0.17 percent compared to OPA. For a 40-unit system, SAMOPA achieves the best cost of 119,369.57 dollars, which is 0.58 percent lower than the best result from OPA. The consistency of SAMOPA's performance is also demonstrated in the Economic Dispatch case with nonlinear valve-point effects and in the CEC 2020 benchmark function, which includes unimodal, hybrid, and composition categories. Statistical validation using the Wilcoxon signed-rank test indicates that SAMOPA exhibits significantly better convergence behavior and solution consistency compared to the comparison method. These results confirm that SAMOPA is an effective and reliable new contribution to solving complex and large-scale economic dispatch problems.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Optimization challenges, particularly Economic Dispatch (ED), aim to minimize the operational costs of power plants while ensuring efficient power distribution [1][2]. ED allocates power among generating units while considering operational and technical constraints [3][4]. Due to its non-linear and complex characteristics, ED remains difficult to solve using gradient-based methods, especially in large-scale systems with multiple decision variables. Consequently, metaheuristic algorithms such as Particle Swarm Optimization (PSO) [5][6][7], Bat Algorithm [8], Grey Wolf Optimizer (GWO) [9][10], and Whale Optimization Algorithm (WOA) [11] have emerged as more flexible and efficient alternatives [12]. Table 1 summarizes various algorithm types, their strengths and weaknesses, and their effectiveness across conventional, artificial intelligence (AI)-based, and hybrid approaches to ED problems.

Table 1: Comparison of Recent Approaches for Economic Dispatch.

Approach Type	Representative Methods	Strengths	Weaknesses	Performance Metrics
Conventional [13]	Lambda Iteration, Newton's Method, Interior Point Method, Quadratic Programming	Simple, efficient for small systems, fast implementation	Not effective for large-scale and nonlinear systems	Computational efficiency, solution accuracy
AI-Based [13]	Artificial Neural Networks, Fuzzy, Deep Learning, Reinforcement Learning	Adaptive, capable of handling complex systems, suitable for real-time use	Prone to overfitting, requires large datasets and parameter tuning	Prediction accuracy, generalization ability
Hybrid [13]	PSO+ANN, Fuzzy+Genetic Algorithm, Metaheuristic Hybrid	Combines the strengths of two methods, provides better solutions	Complex implementation and design, more difficult parameter tuning	Computational time, solution quality
Machine Learning -Based [14]	ML-optimized Droop Control	Minimizes generation cost and power loss simultaneously, no central communication needed	Requires model training, sensitive to training data quality	Generation cost, power loss, system stability

Source: Authors, (2026).

Despite substantial advancements in traditional and AI-based optimization methods, these approaches remain limited in accuracy, computational efficiency, and robustness when addressing nonlinear and large-scale challenges. Hybrid AI–physics approaches, which combine machine learning models such as neural networks with classical optimization techniques like PSO or evolutionary algorithms, have gained significant traction in power system applications [15][16][17]. Although effective in handling non-linearity and system complexity, these hybrid methods often demand significant computational resources. To address this gap, this study proposes the Self-Adaptive Momentum Orca Predation Algorithm (SAMOPA), which introduces a new self-adaptive momentum mechanism into the original Orca Predation Algorithm (OPA) framework. This study is the first to combine adaptive momentum strategies, originally proposed by Polyak (1964), to accelerate convergence in gradient-based methods in OPA to solve ED problems [18][19][20]. SAMOPA improves convergence speed, computational efficiency, and solution stability by dynamically adjusting the movement of search agents based on historical progress. Unlike previous OPA variants that improve solution quality through Levy flight or opposition-based learning but incur higher computational costs [21], SAMOPA achieves better cost reduction and faster execution time with minimal additional complexity. Specifically, empirical results show that SAMOPA reduces total generation costs by up to 0.58%, accelerates convergence time by over 70% compared to OPA, and maintains stable performance in ED scenarios with 6, 15, and 40 unit system.

The main contributions of this paper are as follows:

1. Proposal of SAMOPA, a new OPA variant with a self-adaptive momentum update rule tailored to enhance convergence behavior in solving ED problems.
2. Comprehensive benchmarking against six popular algorithms, OPA, PSO, Bat Algorithm, GWO, WOA, and CMA-ES, on ED test cases of increasing complexity.
3. Quantitative validation of improvements in terms of cost reduction, solution robustness, and computational efficiency.
4. Demonstration of SAMOPA's consistent performance in economic dispatch scenarios with nonlinear valve-point effects, as well as on the CEC 2020 benchmark functions covering unimodal, hybrid, and composition categories.
5. Statistical validation through the Wilcoxon signed-rank test confirms that SAMOPA exhibits significantly better convergence behavior and solution consistency compared to the benchmark algorithms.

The remainder of this paper is structured as follows: Section 2 outlines the research methodology, including the ED problem formulation, the SAMOPA algorithm, and the experimental setup. Section 3 presents the results and comparative analysis, while Section 4 concludes the study and suggests directions for future research.

II. MATERIALS AND METHODS

II.1 SELF ADAPTIVE MOMENTUM ORCA PREDATION ALGORITHM

The original OPA, developed by, is a metaheuristic inspired by the hunting strategies of orca whales, consisting of a chase phase and an assault phase. The algorithm updates the orca's position based on probabilistic rules; however, it is prone to excessive oscillation and premature convergence, especially during early iterations or near the global optimum [22][23]. To address these limitations, SAMOPA introduces a momentum-based enhancement in the chase phase, enabling the search agent to retain directional information from previous iterations. This adjustment improves search stability and promotes more focused exploration [24]. The computational speed of the solution at iteration is defined as equations (1) and (2):

$$v^t_{chase\ 1,i} = m * a\left(dx^t_{best} - F(bM^t + cx^t_i)\right) \quad (1)$$

and

$$v^t_{chase\ 2,i} = m * (ex^t_{best} - x^t_i) \quad (2)$$

Let t denote the number of cycles, $v_{chase\,1,i}^{t}$ represent the chasing speed, M signify the average position of the orca group, x_i^t indicate the position of the orca, the orca with the highest fitness value is marked as x_{best}^{t}, while $a, b,$ and d are random variables within the interval [0,1]. Additionally, e is a random variable within the interval [0,2] and F equals 2. The dynamic modification of m, a momentum component, ensures a balance between exploration and exploitation [25]. Equation (3) articulates this modification:

$$m = m_0 \cdot (1 - \alpha \cdot t) \tag{3}$$

where m_0 represents the initial momentum, valued at 1. The initial momentum value should guarantee optimal exploration in the early phases. α denotes the momentum decay rate of 0.05. The momentum parameter $m_0 = 1$ is employed to augment the algorithm's exploration capacity during the initial phase by leveraging the agent's historical movement data. The decay rate $\alpha = 0.05$ progressively diminishes this influence over time, so facilitating a seamless shift from the exploration to the exploitation phase. This adaptive modification maintains solution diversity in the initial phase and enhances convergence in the final phase. Various combinations of α $(0.01, 0.03, 0.05, \text{and } 0.1)$ and m_0 $(0.5, 1.0, \text{and } 1.5)$ have been evaluated on multiple small-scale economic dispatch problems, with the chosen combination yielding the highest consistent convergence rate and solution quality. This methodology aligns with the DEMON strategy proposed by [26], which demonstrated that progressive momentum decay enhances outcomes by stabilizing parameter updates in the later phase while maintaining flexibility in the initial phase. Adding adaptive momentum to SAMOPA speeds up the search for the best solution, avoids local stagnation, and keeps exploration stable over many iterations. This leads to more accurate and effective solutions, especially for big optimization problems like Economic Dispatch.

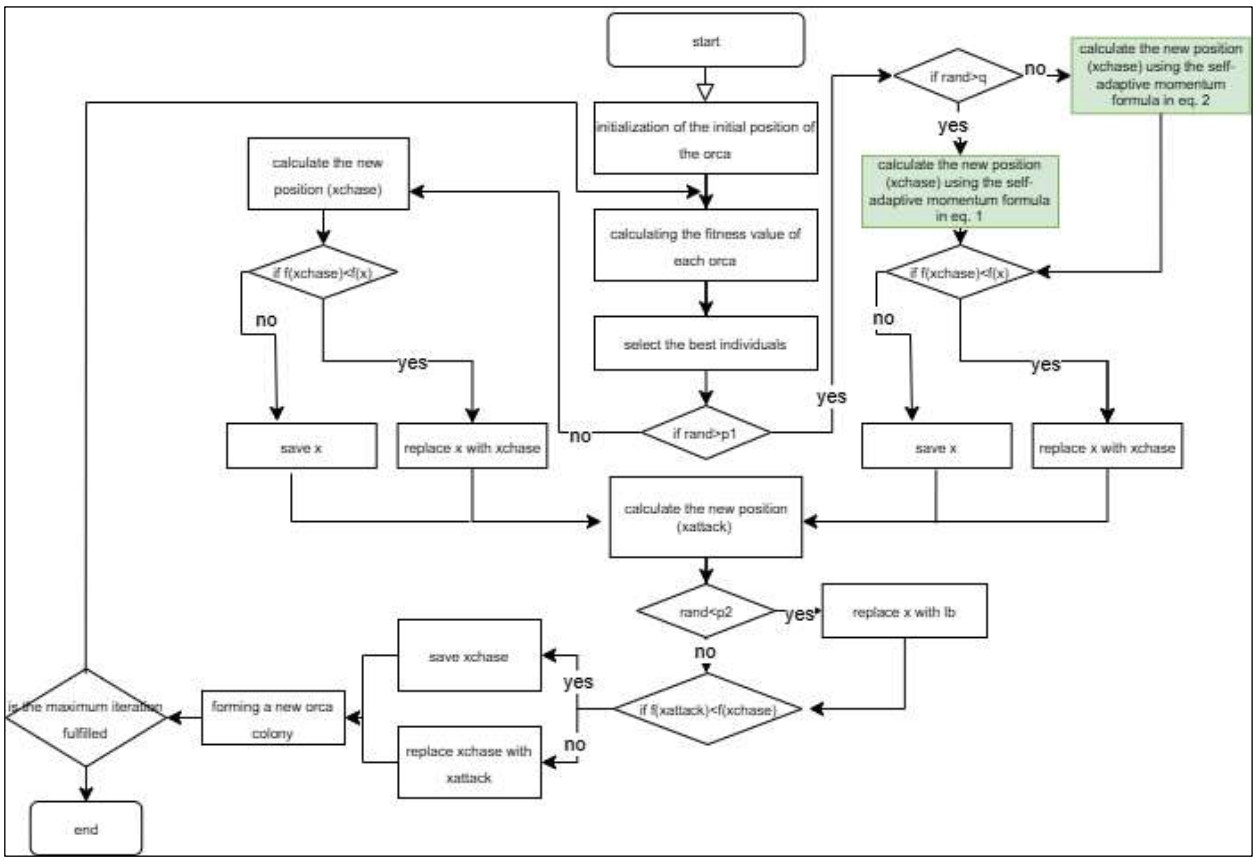


Figure 1: Flowchart of SAMOPA.
Source: Authors, (2026).

The flowchart of SAMOPA in Figure 1 begins with the initialization of the initial position of the orca using a uniform distribution, followed by the calculation of fitness values to determine the best individuals. Based on random values compared to a certain threshold, the system will choose whether to pursue a pursuit strategy or an attack strategy. If the pursuit strategy is chosen, adaptive momentum is used to maintain the direction of change from the previous iteration, making exploration more stable and focused. After evaluation, the new position is saved if it is better than the previous one. If the termination criteria have not been met, the process continues with the formation of a new colony, and iterations continue until the optimal solution is achieved. The combination of these two strategies allows the modified algorithm to achieve more efficient and effective optimization results.

II.2 ECONOMIC DISPATCH PROBLEM FORMULATION

ED is a fundamental optimization problem in power systems that aims to allocate generation output among multiple units in a cost-efficient manner [27]. It is typically formulated as a mathematical model that characterizes the relationship between each unit's power output and its associated operating cost. The total generation cost is commonly modeled as a quadratic function, as defined in Equation (4):

$$F_{Total} = \sum_{i=1}^{n} (a_i P_i^2 + b_i P_i + c_i) \quad (4)$$

In this equation, F_{Total} signifies the total operational cost in currency units per hour, while P_i indicates the output power of the i-th power plant in megawatts (MW). Parameters a_i, b_i, c_i represent the operational cost coefficients for each plant, whereas N denotes the total number of plants involved [13]. The primary challenge in ED is to minimize overall operational costs while adhering to diverse technological and operational restrictions. To ensure physical feasibility, the ED problem must comply with two main types of constraints: equality and inequality constraints. The power balance constraint, presented in Equation (5), ensures that the total generated power meets the system demand:

$$P_D = \sum_{i=1}^{n} P_i \quad (5)$$

where P_D is the overall power demand in megawatts (MW) [28]. In this study, the power balance is enforced using a repair-based approach, where the power output of the first unit is adjusted according to Equation (6):

$$P_1 = P_D - \sum_{i=2}^{n} P_i \quad (6)$$

Should the altered value of P_1 be above the unit's operational threshold, the cost function incurs significant penalties, rendering the solution unviable [29]. Furthermore, each generating unit must function within its minimal ($P_{i,min}$) and maximum ($P_{i,max}$) capacity constraints, as in equation 7.

$$P_{i,min} \leq P_i \leq P_{i,max} \text{ for } i = 1, 2, \ldots, N \quad (7)$$

These restrictions ensure the safety and reliability of system operations. Accordingly, the ED task is to determine an optimal generation schedule that minimizes total operational cost while meeting load requirements and respecting technical constraints. The constraint-handling mechanism, particularly the power balance adjustment, is essential for preserving the feasibility and stability of candidate solutions throughout the optimization process [30].

II.3 EXPERIMENTAL SETUP

This study investigates the ED problem across systems comprising 6, 15, and 40 generating units. Each unit is characterized by minimum ($P_{i,min}$) and maximum ($P_{i,max}$) power generation limits. For instance, the total load demand is set to 1263 MW for the 6-unit case, 2630 MW for the 15-unit case, and 10500 MW for the 40-unit case. The operational cost coefficients (a_i, b_i, c_i) are drawn from a well-established dataset that accurately reflects real-world plant characteristics, as presented by [31]. These simulations enforce both unit-level generation constraints and the global power balance condition as defined in Equation (7). To evaluate the generalization capability of SAMOPA across various optimization landscapes, the CEC 2020 benchmark functions are employed. These functions span four categories: unimodal, basic, hybrid, and composition. Comparative assessments are conducted against the original OPA, as well as other well-known metaheuristic algorithms, including PSO [32][33], GWO [34][35], WOA [36][37], the Bat Algorithm [38][39], and CMA-ES [40][41]. This study employed the Wilcoxon Signed-Rank Test, a non-parametric statistical method, to compare the performance of SAMOPA with other algorithms by assessing the median difference between two matched data sets. Each algorithm underwent testing across 30 independent iterations for each problem, and the convergent outcomes (global minimum) of the two algorithms were juxtaposed.

The null hypothesis (H_0) posits that there is no substantial disparity between the median answers produced by the two algorithms. If the P-value derived from this test is less than 0.05, the null hypothesis is rejected, signifying a substantial difference between the algorithms being tested. The symbols +, -, and = denote the superiority, inferiority, or equivalence of an algorithm, contingent upon a comparison of the ranks of the generated solutions. The parameter configurations of all algorithms in this study were uniformly established to guarantee equitable comparison and dependable replication. The population size was established at 30 for all scenarios, with a maximum of 40 iterations for the 6- and 15-generators cases and 100 iterations for the 40-generators case to ensure sufficient exploration of the broader search space. The parameter settings for each algorithm are modified based on corroborated prior research, including OPA as per Jiang et al.[22], PSO and GWO as outlined by Pokala and Lalitha [42], Bat incorporating enhancements from Meng et al. [43], WOA adhering to the revised version by Chakraborty et al. [44], and CMA-ES reflecting the original configuration by Hansen and Ostermeier [45]. All algorithms were implemented using Python 3.11.12 and run using Google Colab with uniform hardware conditions.

This consistent computing environment ensures that computational time comparisons between methods are fair and reliable. The SAMOPA implementation in Python, along with the entire experimental setup, is available upon request to support transparency and replication in future research. The proposed algorithm, SAMOPA, follows the OPA guidelines for all main settings, except for two key settings: p_1 (the chance of the pursuit phase) and p_2 (the chance of resetting in the attack phase), which are set using a grid search method. This process is carried out on three ED test cases with 6, 15, and 40 generators to find the best combination of settings that results in the lowest cost. The test findings indicate distinct ideal parameter combinations for each scenario: p_1 = 0.9 and p_2 = 0.01 for 6 units, p_1 = 0.9 and p_2 = 0.06 for 15 units, and p_1 = 0.6 and p_2 = 0.04 for 40 generators. The simulation exercise employed fundamental assumptions to streamline the calculation model and enhance the emphasis on the algorithm's performance evaluation. The power load is presumed constant throughout the optimization process, excluding transmission losses and ramp rate constraints. All producing units are presumed to be continuously operational and steady throughout the simulation time.

In the traditional scenario, the generation cost function is presumed to be quadratic and continuous, adhering to the standard economic dispatch model. In the supplementary scenario involving the valve-point impact, the cost function is rendered non-linear and non-convex, accurately representing the actual situations in steam turbines. Physical limitations on the minimum and maximum power of each unit are explicitly considered, and the system power equilibrium is preserved by a repair-based method, by adjusting the power at the initial unit to ensure that total generation meets load demand. All simulations were conducted in 30 independent iterations for each scenario (6, 15, and 40 units) to verify statistical validity.

III. RESULT AND DISCUSSION

III.1 PERFORMANCE OF ECONOMIC DISPATCH PROBLEMS

The performance assessment of SAMOPA in ED with 6, 15 and 40 generators is presented in Table 2.

Table 2: Performance Comparison of the Algorithm on ED Problems.

Algorithm	Units	Best Cost	Mean Cost	Std Dev	Time (s)
SAMOPA	6	**15275.9304**	**15275.9304**	**2.54E-06**	**0.2677**
OPA	6	15275.9306	15275.9327	0.0029	0.9751
PSO	6	15275.9335	15276.7886	2.2008	0.0605
GWO	6	15276.0137	15281.362	10.2684	0.1454
WOA	6	15276.8222	15286.5137	7.5818	0.0388
Bat	6	15283.5699	15306.6403	28.319	0.1179
CMAES	6	15278.6060	15293.0752	10.5748	0.0326
SAMOPA	15	**32256.879**	**32257.767**	**0.646**	**0.266**
OPA	15	32257.415	32258.047	0.606	0.378
PSO	15	32258.844	32369.053	97.605	0.109
GWO	15	32262.852	32417.236	234.646	0.209
WOA	15	32296.006	32452.31	82.416	0.043
Bat	15	32480.021	32733.632	88.709	0.123
CMAES	15	32296.413	32369.321	260.947	0.023
SAMOPA	**40**	**119369.5713**	**120062.3182**	**513.0342**	**0.9378**
OPA	40	120066.9802	121098.2238	742.494	1.4413
PSO	40	122699.5808	133251.8863	5827.9833	0.2936
GWO	40	126436.5593	131538.108	3331.3004	0.872
WOA	40	124377.5908	130962.0752	3550.0292	0.0955
Bat	40	132030.8897	9333342303	2494404695	0.1958
CMAES	40	147602.7932	9333343234.481	3331.3004	0.0577

Source: Authors, (2026).

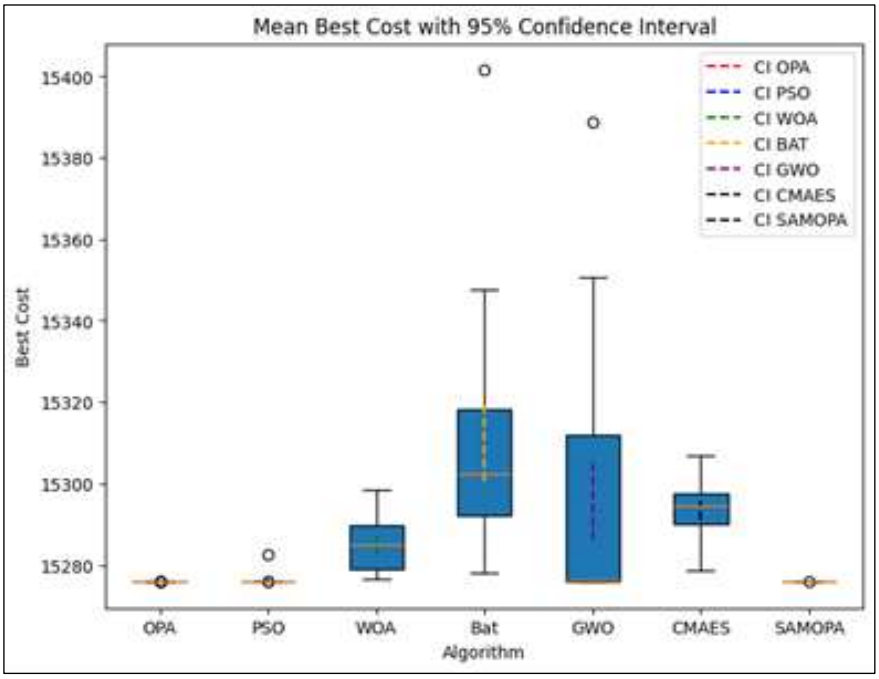


Figure 2: Mean plot with a 95% confidence interval for the algorithms compared across 6 generators.
Source: Authors, (2026).

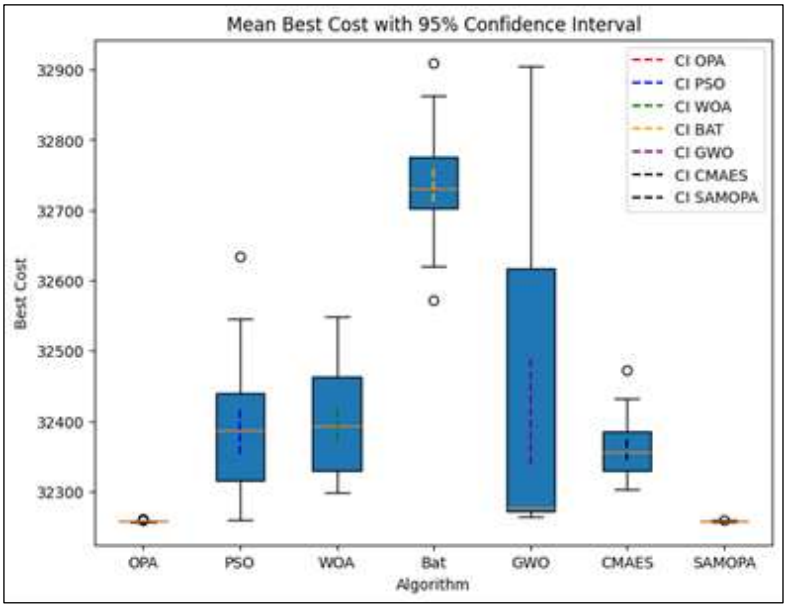


Figure 3: Mean plot with a 95% confidence interval for the algorithms compared across 15 generators.
Source: Authors, (2026).

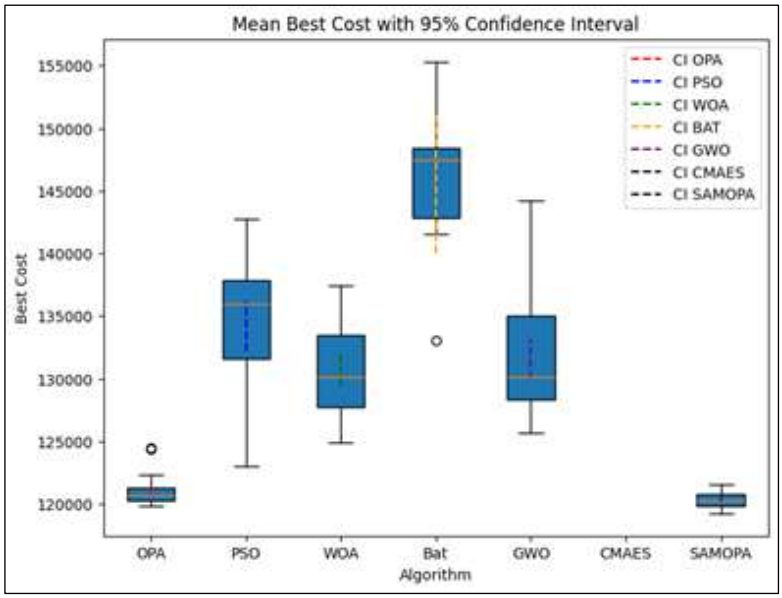


Figure 4: Mean plot with a 95% confidence interval for the algorithms compared across 40 generators.
Source: Authors, (2026).

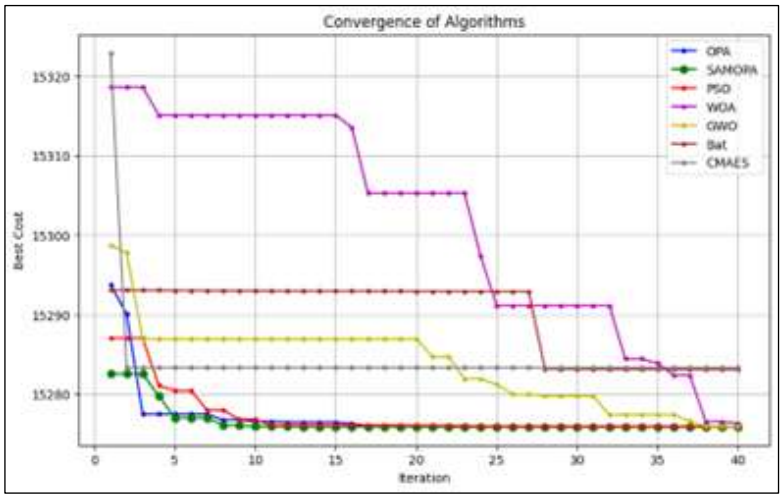


Figure 5: Convergence Curve Comparison of SAMOPA on 6 generators.
Source: Authors, (2026).

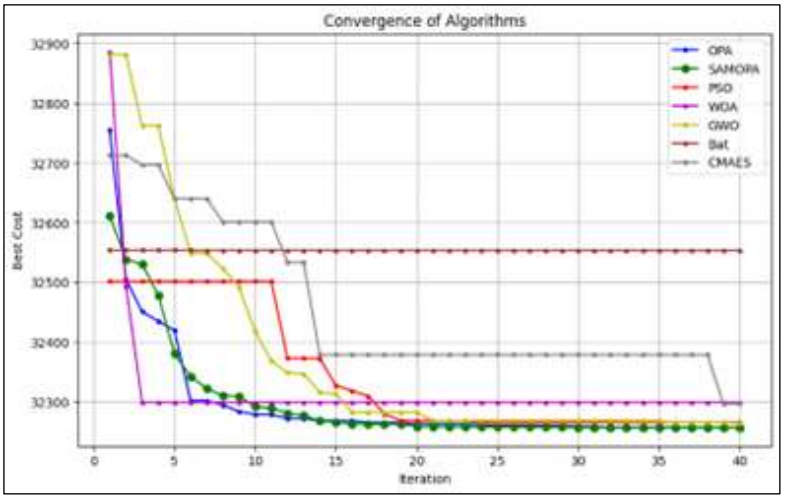


Figure 6: Convergence Curve Comparison of SAMOPA on 15 generators.
Source: Authors, (2026).

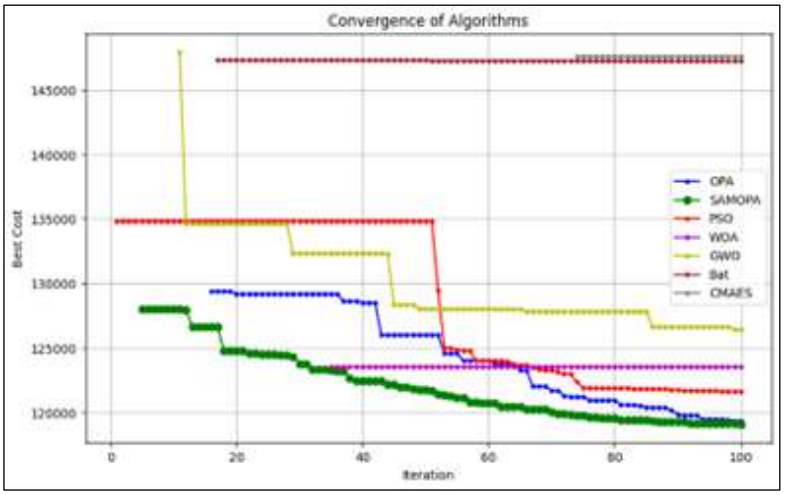


Figure 7: Convergence Curve Comparison of SAMOPA on 40 generators.
Source: Authors, (2026).

Table 2 demonstrates that SAMOPA consistently outperforms other algorithms in economic dispatch scenarios involving 6, 15, and 40 generating units. It achieves the lowest total cost, minimal standard deviation, and competitive computational time, particularly surpassing OPA, PSO, WOA, Bat, GWO, and CMA-ES in terms of solution stability and accuracy. Figures 2–4 further illustrate that SAMOPA consistently yields the lowest mean best cost and the narrowest 95% confidence intervals compared to the benchmark algorithms. The convergence plots in Figures 5–7 reveal that SAMOPA converges more rapidly and maintains greater stability throughout the optimization process. This combination of fast convergence, high accuracy, and robust performance positions SAMOPA as the most reliable method for solving economic dispatch problems across small, medium, and large-scale systems.

III.2 EVALUATION OF NON-CONVEX ED CASES WITH VALVE-POINT EFFECT

To evaluate the robustness of the proposed SAMOPA algorithm under more realistic and complex economic dispatch conditions, further tests were conducted on systems with 6, 15, and 40 generating units incorporating valve-point effects. These effects introduce significant non-linearity and non-convexity into the cost function, closely reflecting practical scenarios in thermal power systems. The performance results for all configurations are presented in Table 3.

Table 3: Performance Comparison of the Algorithm on ED with Valve-Point Effect

Algorithm	Units	Best Cost	Mean Cost	Std Dev	Time (s)
SAMOPA	6	**14326.02**	**14334.26**	**18.41**	0.3788
OPA	6	14326.03	14337.85	21.32	0.5266
PSO	6	14326.09	14389.24	116.65	0.0711
GWO	6	14326.48	14555.78	316.14	0.1427
WOA	6	14354.51	14584.49	142.71	0.0407
Bat	6	14459.1	14885.84	272.8	0.061
CMAES	6	14670.77	-	-	0.0196
SAMOPA	15	**31545.05**	**31644.18**	**72.35**	0.4041
OPA	15	31577.99	31707.11	91.33	0.7249
PSO	15	31826.78	32217.21	265.74	0.1142

Algorithm	Units	Best Cost	Mean Cost	Std Dev	Time (s)
GWO	15	31597.95	32086.14	612.01	0.2621
WOA	15	32144.4	32734.33	266.19	0.0427
Bat	15	32654.2	33252.61	253.83	0.0414
CMAES	15	32598.61	-	612.01	0.0198
SAMOPA	40	116569	118143.2	670.503	1.19
OPA	40	117275.7	119195.6	838.296	1.55
PSO	40	126451.8	9.67E+09	1.8E+09	0.19
GWO	40	-	-	-	-
WOA	40	124307.3	131762	4126.151	0.1
Bat	40	131734.5	-	-	-
CMAES	40	-	-	-	-

Source: Authors, (2026).

The results presented in Table 3 indicate that SAMOPA consistently outperformed all comparative algorithms across the 6, 15, and 40 unit ED cases with valve-point effects. It achieved the lowest generation costs, the smallest standard deviations, and competitive computational times, underscoring its robustness and reliability in addressing non-convex ED problems. OPA ranked second, producing results close to SAMOPA but with noticeably longer computation times. PSO, GWO, and WOA demonstrated moderate effectiveness on smaller systems but showed degraded accuracy and consistency as system size increased. In contrast, Bat and CMA-ES failed to generate valid or stable solutions for the 40-unit case. Overall, SAMOPA exhibited superior accuracy, stability, and computational efficiency, confirming its suitability for solving large-scale, non-convex ED problems with realistic system complexities.

III.3 BENCHMARK EVALUATION USING CEC 2020

The CEC 2020 benchmark suite was selected for its ability to capture key characteristics of the ED problem, including non-linearity, multi-modality, and constraint handling. Its complex function landscape, particularly in hybrid and composition categories, closely reflects real-world challenges such as valve-point effects and operational limits in power systems. The performance of SAMOPA on these benchmark functions is summarized in Table 4.

Table 4: Performance Comparison of the SAMOPA Algorithm on CEC 2020 function.

Algorithm		Unimodal Function	Basic Function	Hybrid Function	Composition Function
PSO	**Best**	8.59E+04	1.10E+03	2.33E+03	3.46E+03
	Mean	4.16E+08	1.61E+03	2.33E+03	3.46E+03
	Std	5.59E+08	1.17E+03	3.53E-01	1.52E+00
	Time	1.20E-01	1.30E-01	1.40E-01	2.30E-01
Bat	**Best**	9.30E+09	1.21E+04	2.39E+03	3.53E+03
	Mean	2.12E+10	1.96E+04	2.85E+03	4.85E+03
	Std	6.31E+09	5.14E+03	2.55E+02	9.77E+02
	Time	8.00E-02	9.00E-02	2.30E-01	2.10E-01
WOA	**Best**	5.68E+09	6.64E+03	2.36E+03	3.46E+03
	Mean	1.34E+10	1.96E+04	2.42E+03	3.60E+03
	Std	3.71E+09	4.88E+03	4.60E+01	1.59E+02
	Time	7.00E-02	1.30E-01	1.00E-01	1.60E-01
GWO	**Best**	4.55E+05	1.10E+03	2.33E+03	3.46E+03
	Mean	6.51E+07	1.16E+03	2.33E+03	3.46E+03
	Std	1.09E+08	9.37E+01	1.02E+00	1.85E+00
	Time	3.50E-01	3.90E-01	1.80E-01	2.20E-01
CMA-ES	**Best**	2.06E+08	1.18E+03	2.33E+03	3.46E+03
	Mean	1.32E+09	1.89E+03	2.36E+03	3.49E+03
	Std	2.08E+07	1.19E+02	7.73E-01	2.79E+00
	Time	4.39E-02	7.01E-02	8.47E-02	1.21E-01
OPA	**Best**	1.94E+02	1.10E+03	2.33E+03	3.46E+03
	Mean	2.60E+03	1.10E+03	2.33E+03	3.46E+03
	Std	3.62E+03	2.00E-02	9.00E-02	4.52E-01
	Time	6.50E-01	6.50E-01	6.90E-01	5.30E-01
SAMOPA	**Best**	1.00E+02	1.10E+03	2.33E+03	3.46E+03
	Mean	1.01E+02	1.10E+03	2.33E+03	3.46E+03
	Std	2.19E+00	1.14E-06	5.00E-02	2.43E-01
	Time	4.90E-01	5.40E-01	5.90E-01	7.50E-01

Source: Authors, (2026).

Table 4 presents the benchmark evaluation results on the CEC 2020 test suite, demonstrating that SAMOPA consistently outperforms all comparative algorithms across all function categories—unimodal, basic, hybrid, and composition. It achieved the best objective values, the lowest mean costs, and the smallest standard deviations, reflecting superior accuracy, stability, and convergence reliability. For example, SAMOPA obtained an optimal value of 1.00E+02 in the unimodal function with a mean of 1.01E+02 and a minimal deviation of 2.19E+00, significantly outperforming OPA, PSO, Bat, and WOA. Similarly, in the basic function category, SAMOPA maintained an average of 1.10E+03 with an extremely low variance of 1.14E−06, indicating consistent convergence.

Its performance remained robust in hybrid and composition functions, with variances of just 5.00E−02 and 2.43E−01, whereas other algorithms such as CMA-ES, GWO, and WOA exhibited greater variability. Although SAMOPA incurred slightly higher computation times than some competitors, the trade-off is justified by its consistent optimization quality. These results confirm that the adaptive momentum mechanism significantly enhances the original OPA framework. A series of Wilcoxon signed-rank tests were conducted on four types of benchmark functions, Unimodal, Basic, Hybrid, and Composition, to statistically confirm that the SAMOPA algorithm is better than other algorithms. We conducted the tests in pairs, comparing SAMOPA with six comparative algorithms: OPA, PSO, WOA, Bat, GWO, and CMA-ES. The findings of this statistical test are summarized in Table 5 below.

Table 5: Statistical Result using Wilcoxon Signed Rank.

	SAMOPA vs OPA				**SAMOPA vs PSO**			
Problem	**P-value**	**T+**	**T−**	**Winner**	**P-value**	**T+**	**T−**	**Winner**
Unimodal Function	1.86E-09	0	1E+05	+	1.86E-09	0	9E+09	+
Basic Function	1.86E-09	0	0	+	1.86E-09	0	42879	+
Hybrid Function	1.05E-01	2	0	=	3.05E-05	0	20	+
Composition Function	2.89E-01	6	2	=	1.23E-04	2	46	+
	SAMOPA vs WOA				**SAMOPA vs Bat**			
Problem	**P-value**	**T+**	**T−**	**Winner**	**P-value**	**T+**	**T−**	**Winner**
Unimodal Function	1.86E-09	0	4E+11	+	1.86E-09	0	4E+11	+
Basic Function	1.86E-09	0	5E+05	+	1.86E-09	0	5E+05	+
Hybrid Function	1.86E-09	0	2993	+	1.86E-09	0	54564	+
Composition Function	1.86E-09	0	6303	+	1.86E-09	0	36461	+
	SAMOPA vs GWO				**SAMOPA vs CMAES**			
Problem	**P-value**	**T+**	**T−**	**Winner**	**P-value**	**T+**	**T−**	**Winner**
Unimodal Function	1.86E-09	0	2E+09	+	1.86E-09	0	3E+10	+
Basic Function	1.86E-09	0	43	+	1.86E-09	0	26256	+
Hybrid Function	1.86E-09	0	48	+	1.86E-09	0	786	+
Composition Function	1.86E-09	0	87	+	1.86E-09	0	1340	+

Source: Authors, (2026).

Table 5 shows the results of the Wilcoxon Signed-Rank Test, indicating that SAMOPA consistently performs better than other algorithms like OPA, PSO, WOA, Bat, GWO, and CMAES on various test functions, including Unimodal, Basic, Hybrid, and Composition. In nearly all comparisons, minimal P-values (<0.05) signify substantial differences, demonstrating that SAMOPA is superior in producing optimal solutions. The '+' symbol in all comparisons shows that SAMOPA usually does better than other algorithms in terms of solution quality and convergence; however, in the Hybrid and Composition functions, the higher P-values indicate that there is no significant difference between SAMOPA and the other algorithms. SAMOPA consistently outperformed its competitors in most comparisons, showcasing its superiority in efficiency and performance throughout the studied optimization functions. SAMOPA demonstrates strong performance across all test scenarios, including both standard and non-convex economic dispatch problems, validating its capability to consistently achieve global optimal solutions. This success is primarily attributed to the adaptive momentum mechanism, which dynamically balances exploration and exploitation by adjusting momentum over iterations.

Unlike fixed momentum approaches, SAMOPA gradually reduces momentum to encourage broad exploration in early stages and focused exploitation in later phases, enhancing both convergence stability and solution quality, as evidenced in Figures 4–6 and Tables 2–3. Although SAMOPA incurs slightly longer computation times than PSO, WOA, and GWO, it consistently delivers more accurate and stable outcomes while avoiding premature convergence. These findings align with prior studies [18], [46], [47], which highlight the benefits of momentum in stabilizing convergence in metaheuristic frameworks. Compared to the original OPA, SAMOPA significantly improves computational efficiency without compromising accuracy. Nonetheless, the comparatively extended execution time relative to alternative algorithms remains a difficulty for real-time applications, particularly in extensive optimization contexts. To address this constraint, future research may investigate the incorporation of hardware acceleration methods, including GPU-based parallelization, multi-core processing, and distributed computing frameworks.

These methodologies facilitate concurrent assessment of populations, expedite the optimization process, and enhance the scalability of algorithms. Moreover, the deployment of cloud infrastructure can enhance the efficiency and real-time responsiveness of SAMOPA. The adaptive momentum mechanism in SAMOPA can be further enhanced through feedback-driven momentum adjustment techniques or dynamic learning rate adaptation, potentially decreasing computing time while maintaining accuracy. Potential hybridization with other metaheuristic algorithms, such as the incorporation of exploitation operators from Particle Swarm Optimization (PSO) or Differential Evolution (DE), may enhance convergence speed and overall solution quality. Finally, while SAMOPA has demonstrated robust empirical convergence, it lacks a mathematical proof of convergence. Consequently, additional theoretical investigations are required to enhance the mathematical foundation and reliability of this methodology over time.

IV. CONCLUSION

This study has proposed and evaluated the SAMOPA as a solution for ED problems in power systems. Experimental results demonstrate that SAMOPA consistently delivers more accurate and reliable solutions than conventional metaheuristic algorithms, even under complex conditions such as valve-point effects. This performance advantage is primarily attributed to the integration of an adaptive momentum mechanism, which effectively balances exploration and exploitation to accelerate convergence.

SAMOPA's strong results across various benchmark scenarios affirm its suitability for real-world power system optimization tasks. Future research will focus on further improving computational efficiency, strengthening the theoretical foundation of the algorithm, and extending its application to ultra-large-scale systems involving 100 or more generating units to evaluate its scalability in more complex and realistic environments.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Vivi Aida Fitria and Arif Nur Afandi.
Methodology: Vivi Aida Fitria and Arif Nur Afandi.
Software: Vivi Aida Fitria.
Validation: Vivi Aida Fitria and Arif Nur Afandi.
Formal Analysis: Arif Nur Afandi.
Visualization: Vivi Aida Fitria.
Writing – Original Draft: Vivi Aida Fitria.
Writing – Review and Editing: Arif Nur Afandi and Aripriharta.
Resources: Arif Nur Afandi.
Supervision: Aripriharta.
Project Administration: Aripriharta.
Approval of the final text: Vivi Aida Fitria, Arif Nur Afandi, and Aripriharta.

VI. REFERENCES

[1] A. B. Kunya, A. S. Abubakar, and S. S. Yusuf, "Review of economic dispatch in multi-area power system: State-of-the-art and future prospective," Electr. Power Syst. Res., vol. 217, p. 109089, Apr. 2023, doi: 10.1016/j.epsr.2022.109089.

[2] S. K. Rangu, P. R. Lolla, K. R. Dhenuvakonda, and A. R. Singh, "Recent trends in power management strategies for optimal operation of distributed energy resources in microgrids: A comprehensive review," Int. J. Energy Res., vol. 44, no. 13, pp. 9889–9911, Oct. 2020, doi: 10.1002/er.5649.

[3] L. Al-Bahrani, M. Seyedmahmoudian, B. Horan, and A. Stojcevski, "Solving the Real Power Limitations in the Dynamic Economic Dispatch of Large-Scale Thermal Power Units under the Effects of Valve-Point Loading and Ramp-Rate Limitations," Sustainability, vol. 13, no. 3, p. 1274, Jan. 2021, doi: 10.3390/su13031274.

[4] A. N. Afandi, I. Fadlika, and A. Andoko, "Comparing Performances of Evolutionary Algorithms on the Emission Dispatch and Economic Dispatch Problem," TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 13, no. 4, p. 1187, Dec. 2015, doi: 10.12928/telkomnika.v13i4.3166.

[5] B. M. Hussein, "Evolutionary algorithm solution for economic dispatch problems," Int. J. Electr. Comput. Eng., vol. 12, no. 3, p. 2963, Jun. 2022, doi: 10.11591/ijece.v12i3.pp2963-2970.

[6] J. Zhang, J. Zhang, F. Zhang, M. Chi, and L. Wan, "An Improved Symbiosis Particle Swarm Optimization for Solving Economic Load Dispatch Problem," J. Electr. Comput. Eng., vol. 2021, pp. 1–11, Jan. 2021, doi: 10.1155/2021/8869477.

[7] A. E. Prasetya and T. Wrahatnolo, "Economic Dispatch Pada Pembangkit Termal Pln Apb Iv Jawa Timur Menggunakan Metode Particle Swarm Optimization (Pso)," J. Tek. Elektro, vol. 9, no. 1, pp. 885–892, 2020, doi: https://doi.org/10.26740/jte.v9n1.p%25p.

[8] F. Tariq, S. Alelyani, G. Abbas, A. Qahmash, and M. R. Hussain, "Solving Renewables-Integrated Economic Load Dispatch Problem by Variant of Metaheuristic Bat-Inspired Algorithm," Energies, vol. 13, no. 23, p. 6225, Nov. 2020, doi: 10.3390/en13236225.

[9] S. Hosseini-Hemati, S. Derafshi Beigvand, H. Abdi, and A. Rastgou, "Society-based Grey Wolf Optimizer for large scale Combined Heat and Power Economic Dispatch problem considering power losses," Appl. Soft Comput., vol. 117, p. 108351, Mar. 2022, doi: 10.1016/j.asoc.2021.108351.

[10] T. Jayabarathi, T. Raghunathan, B. R. Adarsh, and P. N. Suganthan, "Economic dispatch using hybrid grey wolf optimizer," Energy, vol. 111, pp. 630–641, Sep. 2016, doi: 10.1016/j.energy.2016.05.105.

[11] H. J. Touma, "Study of The Economic Dispatch Problem on IEEE 30-Bus System using Whale Optimization Algorithm," Int. J. Eng. Technol. Sci., vol. 3, no. 1, pp. 11–18, Jun. 2016, doi: 10.15282/ijets.5.2016.1.2.1041.

[12] K. E. Fahim, L. C. De Silva, F. Hussain, and H. Yassin, "A State-of-the-Art Review on Optimization Methods and Techniques for Economic Load Dispatch with Photovoltaic Systems: Progress, Challenges, and Recommendations," Sustainability, vol. 15, no. 15, p. 11837, Aug. 2023, doi: 10.3390/su151511837.

[13] F. Marzbani and A. Abdelfatah, "Economic Dispatch Optimization Strategies and Problem Formulation: A Comprehensive Review," Energies, vol. 17, no. 3, p. 550, Jan. 2024, doi: 10.3390/en17030550.

[14] Y. Saeidinia, M. Arabshahi, M. Aminirad, and M. Shafie-khah, "Enhancing DC microgrid performance through machine learning-optimized droop control," IET Gener. Transm. Distrib., vol. 18, no. 9, pp. 1919–1934, May 2024, doi: 10.1049/gtd2.13169.

[15] A. A. Movassagh et al., "Artificial neural networks training algorithm integrating invasive weed optimization with differential evolutionary model," J. Ambient Intell. Humaniz. Comput., vol. 14, no. 5, pp. 6017–6025, May 2023, doi: 10.1007/s12652-020-02623-6.

[16] R. L. Rose, B. D. A. Selvi, and R. L. R. Singh, "Development of Hybrid Algorithm Based on PSO and NN to Solve Economic Emission Dispatch Problem," Circuits Syst., vol. 07, no. 09, pp. 2323–2331, 2016, doi: 10.4236/cs.2016.79202.

[17] S. Tsegaye and G. Bekele, "Optimal Generation Dispatch of Ethiopian Power System Using Hybrid Genetic Algorithm-Hopfield Neural Network," EAI Endorsed Trans. Energy Web, p. 170673, Jul. 2018, doi: 10.4108/eai.13-8-2021.170673.

[18] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," 30th Int. Conf. Mach. Learn. ICML 2013, no. PART 3, pp. 2176–2184, 2013.

[19] H. Sun et al., "An Improved Reacceleration Optimization Algorithm Based on the Momentum Method for Image Recognition," Mathematics, vol. 12, no. 11, p. 1759, Jun. 2024, doi: 10.3390/math12111759.

[20] Y. Li et al., "Q-DADAM: A Quantized Distributed Online Optimization Algorithm With Adaptive Momentum," IEEE Trans. Control Netw. Syst., pp. 1–12, 2025, doi: 10.1109/TCNS.2025.3526555.

[21] E. H. Houssein, M. R. Saad, A. A. Ali, and H. Shaban, "Multiple Strategies Boosted Orca Predation Algorithm for Engineering Optimization Problems," Int. J. Comput. Intell. Syst., vol. 16, no. 1, p. 67, Apr. 2023, doi: 10.1007/s44196-023-00249-y.

[22] Y. Jiang, Q. Wu, S. Zhu, and L. Zhang, "Orca predation algorithm: A novel bio-inspired algorithm for global optimization problems," Expert Syst. Appl., vol. 188, p. 116026, Feb. 2022, doi: 10.1016/j.eswa.2021.116026.

[23] V. A. Fitria, A. N. Afandi, and Aripriharta, "Exploring the Orca Predation Algorithm for Economic Dispatch Optimization in Power Systems," BenchCouncil Trans. Benchmarks, Stand. Eval., p. 100187, Jan. 2025, doi: 10.1016/j.tbench.2024.100187.

[24] J. Chen and Y. Yuan, "Adaptive Latent Factor Analysis via Generalized Momentum-Incorporated Particle Swarm Optimization," in 2022 IEEE International Conference on Networking, Sensing and Control (ICNSC), IEEE, Dec. 2022, pp. 1–6. doi: 10.1109/ICNSC55942.2022.10004062.

[25] B. H. Nguyen, B. Xue, P. Andreae, and M. Zhang, "A New Binary Particle Swarm Optimization Approach: Momentum and Dynamic Balance Between Exploration and Exploitation," IEEE Trans. Cybern., vol. 51, no. 2, pp. 589–603, Feb. 2021, doi: 10.1109/TCYB.2019.2944141.

[26] J. Chen, C. Wolfe, Z. Li, and A. Kyrillidis, "Demon: Improved Neural Network Training With Momentum Decay," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, May 2022, pp. 3958–3962. doi: 10.1109/ICASSP43922.2022.9746839.

[27] R. Ponciroli, N. E. Stauff, J. Ramsey, F. Ganda, and R. B. Vilim, "An Improved Genetic Algorithm Approach to the Unit Commitment/Economic Dispatch Problem," IEEE Trans. Power Syst., vol. 35, no. 5, pp. 4005–4013, Sep. 2020, doi: 10.1109/TPWRS.2020.2986710.

[28] M. Dashtdar et al., "Solving the environmental/economic dispatch problem using the hybrid FA-GA multi-objective algorithm," Energy Reports, vol. 8, pp. 13766–13779, Nov. 2022, doi: 10.1016/j.egyr.2022.10.054.

[29] H. Suyono, E. Subekti, H. Purnomo, T. Nurwati, and R. N. Hasanah, "Economic Dispatch of 500 kV Java-Bali Power System using Hybrid Particle Swarm-Ant Colony Optimization Method," in 2020 12th International Conference on Electrical Engineering (ICEENG), IEEE, Jul. 2020, pp. 5–10. doi: 10.1109/ICEENG45378.2020.9171771.

[30] P. Yi, Y. Hong, and F. Liu, "Initialization-free distributed algorithms for optimal resource allocation with feasibility constraints and application to economic dispatch of power systems," Automatica, vol. 74, pp. 259–269, Dec. 2016, doi: 10.1016/j.automatica.2016.08.007.

[31] H. Hardiansyah, "A Modified Particle Swarm Optimization Technique for Economic Load Dispatch with Valve-Point Effect," Int. J. Intell. Syst. Appl., vol. 5, no. 7, pp. 32–41, Jun. 2013, doi: 10.5815/ijisa.2013.07.05.

[32] A. B. Putra Utama, A. P. Wibawa, M. Muladi, and A. Nafalski, "PSO based Hyperparameter tuning of CNN Multivariate Time- Series Analysis," J. Online Inform., vol. 7, no. 2, pp. 193–202, Dec. 2022, doi: 10.15575/join.v7i2.858.

[33] W. M. A. Wan Abdul Razak, N. Rosmin, A. H. Musta'amal, S. M. Hussin, D. Mat Said, and Aripriharta, "Power loss minimization by optimal allocation and sizing of STATCOM via particle swarm optimization," E3S Web Conf., vol. 516, p. 05003, Apr. 2024, doi: 10.1051/e3sconf/202451605003.

[34] Q. A. Sias, I. Fadlika, I. D. Wahyono, and A. Nur Afandi, "Quasi Z-Source Inverter as MPPT on Renewable Energy using Grey Wolf Technique," in 2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), IEEE, Oct. 2018, pp. 362–366. doi: 10.1109/EECSI.2018.8752686.

[35] L. I. Wong, M. H. Sulaiman, M. R. Mohamed, and M. S. Hong, "Grey Wolf Optimizer for solving economic dispatch problems," in 2014 IEEE International Conference on Power and Energy (PECon), IEEE, Dec. 2014, pp. 150–154. doi: 10.1109/PECON.2014.7062431.

[36] S. Chakraborty, S. Sharma, A. K. Saha, and A. Saha, "A novel improved whale optimization algorithm to solve numerical optimization and real-world applications," Artif. Intell. Rev., vol. 55, no. 6, pp. 4605–4716, Aug. 2022, doi: 10.1007/s10462-021-10114-z.

[37] M. H. Nadimi-Shahraki, H. Zamani, Z. Asghari Varzaneh, and S. Mirjalili, "A Systematic Review of the Whale Optimization Algorithm: Theoretical Foundation, Improvements, and Hybridizations," Arch. Comput. Methods Eng., vol. 30, no. 7, pp. 4113–4159, Sep. 2023, doi: 10.1007/s11831-023-09928-7.

[38] C. A. Diebold, A. Salles, and C. F. Moss, "Adaptive Echolocation and Flight Behaviors in Bats Can Inspire Technology Innovations for Sonar Tracking and Interception," Sensors, vol. 20, no. 10, p. 2958, May 2020, doi: 10.3390/s20102958.

[39] W. Younas et al., "Improving Convergence Speed of Bat Algorithm Using Multiple Pulse Emissions along Multiple Directions," Sensors, vol. 22, no. 23, p. 9513, Dec. 2022, doi: 10.3390/s22239513.

[40] H.-G. Beyer and B. Sendhoff, "Simplify Your Covariance Matrix Adaptation Evolution Strategy," IEEE Trans. Evol. Comput., vol. 21, no. 5, pp. 746–759, Oct. 2017, doi: 10.1109/TEVC.2017.2680320.

[41] Q. Niu, L. Wang, and M. You, "A Covariance Matrix Adaptation Evolution Strategy Based on Cooperative Co-Evolutionary Framework Using Delta Grouping for Large-Scale Dynamic Economic Dispatch," in 2020 The 3rd International Conference on Machine Learning and Machine Intelligence, New York, NY, USA: ACM, Sep. 2020, pp. 122–126. doi: 10.1145/3426826.3426849.

[42] J. Pokala and B. Lalitha, "A Novel Intrusion Detection System for RPL Based IoT Networks with Bio-Inspired Feature Selection and Ensemble Classifier." Apr. 28, 2021. doi: 10.21203/rs.3.rs-442429/v1.

[43] X.-B. Meng, X. Z. Gao, Y. Liu, and H. Zhang, "A novel bat algorithm with habitat selection and Doppler effect in echoes for optimization," Expert Syst. Appl., vol. 42, no. 17–18, pp. 6350–6364, Oct. 2015, doi: 10.1016/j.eswa.2015.04.026.

[44] S. Chakraborty, A. Kumar Saha, S. Sharma, S. Mirjalili, and R. Chakraborty, “A novel enhanced whale optimization algorithm for global optimization,” Comput. Ind. Eng., vol. 153, p. 107086, Mar. 2021, doi: 10.1016/j.cie.2020.107086.

[45] N. Hansen and A. Ostermeier, “Completely Derandomized Self-Adaptation in Evolution Strategies,” Evol. Comput., vol. 9, no. 2, pp. 159–195, Jun. 2001, doi: 10.1162/106365601750190398.

[46] P. Dhal and C. Azad, “Hybrid momentum accelerated bat algorithm with GWO based optimization approach for spam classification,” Multimed. Tools Appl., vol. 83, no. 9, pp. 26929–26969, Sep. 2023, doi: 10.1007/s11042-023-16448-w.

[47] M. M. Nilsen, A. S. Stordal, R. J. Lorentzen, P. N. Raanes, and K. S. Eikrem, “Accelerated ensemble optimization using momentum methods,” Sci. Rep., vol. 14, no. 1, p. 25336, Oct. 2024, doi: 10.1038/s41598-024-76916-7.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 277-286. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2892

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

OPTIMIZATION OF CONVOLUTIONAL NEURAL NETWORKS THROUGH SOFT VOTING ENSEMBLE AND FINE-TUNED APPROACH TO IMAGE CLASSIFICATION

Johan Wijaya Kusuma*[1], Julian Supardi[2]

[1, 2]Department of Informatics Engineering, Sriwijaya University, P.O BOX 30129, Palembang, Indonesia.

[1]https://orcid.org/0009-0004-7906-4097, [2]https://orcid.org/0000-0002-5836-9236

Email: *kusumajohanwijaya@gmail.com, julian@unsri.ac.id

ARTICLE INFO

Article History
Received: November 5, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
ResNet,
EfficientNet,
DenseNet,
Ensemble Learning,
Image Classification.

ABSTRACT

This study introduces an optimization method to enhance image classification accuracy using a Soft Voting Ensemble of fine-tuned convolutional neural networks (CNNs). The goal is to evaluate whether combining multiple CNN architectures can outperform individual models. Three pre-trained networks DenseNet, ResNet, and EfficientNet were fine-tuned through transfer learning on a Kaggle image dataset. To improve generalization, data augmentation techniques such as random rotation, flipping, and zooming were applied. After fine-tuning, the models were integrated using a soft voting strategy that averages prediction probabilities to determine the final class. Performance was tested on 94 unseen images, fully separated from the training data. The individual accuracies were 91.5% for DenseNet, 87.2% for ResNet, and 93.6% for EfficientNet. The proposed ensemble achieved the highest accuracy of 94.7%, with precision, recall, and F1-score all reaching 0.95. These findings indicate that the ensemble approach successfully combines the strengths of different CNNs, reduces classification errors, and increases model robustness. Overall, the soft voting ensemble provides a reliable, scalable, and effective solution for improving CNN-based image classification, especially when dealing with limited or diverse datasets.

Copyright ©2025 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

In recent years, the fields of computer vision and deep learning have experienced significant advancement, driven by the rapid growth of computational power and the availability of large-scale annotated datasets. Deep learning models, particularly convolutional neural networks (CNNs), have become the backbone of modern image classification systems due to their ability to automatically learn hierarchical feature representations from raw pixel data. Unlike traditional image processing methods that depend on handcrafted features, CNNs can effectively capture spatial dependencies, textures, and semantic structures directly from the data. This capability has enabled breakthroughs in diverse domains such as medical imaging, autonomous driving, surveillance, agriculture, and industrial automation, where visual understanding plays a crucial role in decision-making processes [1].

The fundamental strength of CNNs lies in their hierarchical feature extraction capability, where low-level filters capture edges and textures, while deeper layers represent complex and abstract concepts. Over the past decade, numerous CNN architectures have been proposed to improve model depth, efficiency, and generalization. Among the most influential architectures are ResNet, DenseNet, and EfficientNet, which have become benchmarks in computer vision research and applications. ResNet introduced residual connections to address the vanishing gradient problem and enable the training of very deep networks [2]. DenseNet, on the other hand, employs dense connections among layers to promote feature reuse and gradient flow efficiency, leading to compact yet powerful models [3].

EfficientNet introduced a compound scaling technique that systematically balances network depth, width, and input resolution to achieve state-of-the-art accuracy with fewer parameters [4]. Despite their impressive performance, single CNN architectures still face several challenges. In practice, the performance of a CNN often depends on the quality, size, and diversity of the training dataset. When data is limited, noisy, or exhibits high intra-class variability, models tend to overfit or become biased toward dominant features. Moreover, a single model might excel at identifying certain visual patterns but perform poorly on others, depending on the nature of its learned representations. This model-specific bias often limits generalization, especially when deployed in real-world scenarios with unseen data distributions. Therefore, improving classification stability and robustness remains an open challenge in deep learning–based vision systems [5]. One promising direction to overcome these limitations is ensemble learning. Ensemble methods combine the predictions of multiple models to produce a more reliable and accurate output than any individual model could achieve alone. In traditional machine learning, ensemble strategies such as bagging, boosting, and stacking have demonstrated consistent improvements in performance and stability. In deep learning, similar principles can be applied by combining the outputs of multiple neural networks trained with different initializations, architectures, or subsets of data. Ensemble learning leverages the diversity among models both in terms of structure and learned representations to reduce variance and increase predictive robustness [6].

Among various ensemble strategies, soft voting has emerged as a particularly simple yet effective approach. In soft voting, the probabilistic predictions (output probabilities) from several base classifiers are averaged or weighted to determine the final class prediction. Unlike hard voting, which only considers the majority vote of discrete class labels, soft voting retains the confidence level of each model's prediction, allowing the ensemble to make more nuanced decisions. This approach is especially beneficial when combining heterogeneous deep models, as their probability distributions can complement one another, leading to improved accuracy and smoother decision boundaries [6]. In this study, a Soft Voting Ensemble approach is proposed to enhance image classification performance by combining multiple fine-tuned CNN models. Specifically, three high-performing architectures DenseNet, ResNet, and EfficientNet are fine-tuned through transfer learning on a custom image dataset obtained from Kaggle. Transfer learning allows leveraging pre-trained weights from large-scale datasets such as ImageNet, accelerating convergence and improving generalization even with limited target data. To further increase model robustness, data augmentation techniques are applied during training. Augmentation operations such as random rotation, flipping, scaling, and zooming help to artificially expand the training data and expose the models to diverse visual variations.

This not only mitigates overfitting but also improves the adaptability of CNNs to unseen real-world conditions [7]. The motivation behind using multiple CNNs in an ensemble lies in their complementary feature extraction capabilities. Each architecture has unique design principles that influence how it interprets image information. For instance, ResNet's skip connections enable efficient learning of residual mappings, making it particularly effective for recognizing global patterns and coarse structures. DenseNet's densely connected layers enhance information flow and encourage feature reuse, which benefits texture-rich or detailed images. Meanwhile, EfficientNet's compound scaling provides a balanced architecture that captures both fine and coarse details efficiently. By combining these models using a soft voting mechanism, the ensemble benefits from the diverse representation power of each CNN, leading to improved classification stability and accuracy across different image types. In practical applications, ensemble methods also provide an additional advantage: model uncertainty reduction.

Since deep networks are often sensitive to initialization and data variations, a single model might produce inconsistent predictions when exposed to slightly altered inputs. The ensemble approach mitigates this by averaging multiple predictions, which statistically reduces variance and enhances decision confidence. This makes soft voting particularly suitable for safety-critical applications such as medical diagnosis or industrial inspection, where prediction reliability is paramount [8]. This research aims to achieve three main objectives. First, to investigate the effectiveness of the soft voting ensemble in improving image classification accuracy compared to individual fine-tuned CNN models. Second, to evaluate the impact of transfer learning and data augmentation on model performance and generalization. Third, to analyze the complementary behavior among DenseNet, ResNet, and EfficientNet when combined under a unified probabilistic voting scheme. Through these objectives, this study contributes to the understanding of how ensemble deep learning can be leveraged for more robust and scalable visual recognition systems, particularly in domains where labeled data are scarce or heterogeneous.

II. MATERIALS AND METHODS

II.1 DATASET

The dataset utilized in this study comprises a single image classification dataset focusing on *Pempek*, a traditional Indonesian delicacy originating from Palembang. Figure 1 illustrates representative examples from the *Pempek* image classification dataset. The dataset was obtained from the publicly available Kaggle repository and consists of five distinct classes representing various types of Pempek: Pempek Lenjer, Pempek Kapal Selam, Pempek Adaan, Pempek Keriting, and Pempek Kulit. The images within the dataset exhibit considerable diversity in terms of lighting conditions, backgrounds, and camera orientations, thereby ensuring a comprehensive representation of real-world variations. Such heterogeneity is essential for enhancing the generalization capability of the proposed model.



Figure 1: Example Images.
Source: Authors, (2026).

To strengthen the dataset and mitigate the risk of model overfitting, several data augmentation techniques were employed, including random rotation, horizontal and vertical flipping, cropping, brightness adjustment, and image resizing. These transformations not only increased the effective dataset size but also improved the robustness of the model when exposed to unseen data. Prior to model training, a series of data preprocessing steps were conducted to standardize and optimize the dataset. These include:

- Resizing with Black Padding: All images were resized to a uniform resolution of 224 × 224 pixels. To preserve the original aspect ratio and prevent geometric distortion, any empty regions resulting from resizing were filled with black pixels.
- Data Augmentation: Augmentation operations were systematically applied to the training data to increase variability and improve the model's resilience to environmental and perceptual changes.
- Normalization: Each pixel value was normalized to a range of [0, 1] by dividing by 255. This normalization process ensures numerical stability and accelerates model convergence during training.

Initially, the dataset contained 485 original images. These images were randomly divided into 391 training samples (approximately 80%) and 94 testing samples (approximately 20%). To further improve the robustness of the model and mitigate overfitting, data augmentation was applied exclusively to the training set. Augmentation techniques included random rotation, horizontal and vertical flipping, cropping, brightness adjustment, and resizing. As a result, the number of training images increased from 391 to 2,904 augmented samples, thereby expanding the dataset's variability and improving the model's resilience to real-world conditions. The final dataset comprised a total of 2,904 images across five categories, as summarized in Table 1.

Table 1: Dataset.

Dataset	Source	Classes	Original Images	Test Images	Train Images	After Augmentation
Pempek Food Picture	Kaggle	5	485	94	391	2904

Source: Authors, (2026).

II.2 PROPOSED METHODS

II.2.1 CNN RESNET

ResNet (Residual Network) introduces the principle of residual learning to effectively mitigate the vanishing gradient problem that typically arises in deep convolutional neural networks [9]. By integrating shortcut or skip connections that bypass one or more layers, ResNet facilitates the training of very deep models while maintaining stability and accuracy. In this study, the ResNet50 architecture is employed, consisting of 50 layers organized into a series of convolutional and identity blocks that incorporate batch normalization and ReLU activation functions. Each residual block employs identity mappings and 1×1 convolutions to adjust feature dimensions, enabling efficient hierarchical feature extraction. ResNet50 offers a favorable balance between model complexity and classification accuracy, outperforming shallower variants such as ResNet34 in capturing high-level semantic features while remaining computationally practical for fine-tuning tasks [10]. The network is initialized using pre-trained ImageNet weights, which accelerates convergence and enhances the model's generalization performance within the transfer learning framework.

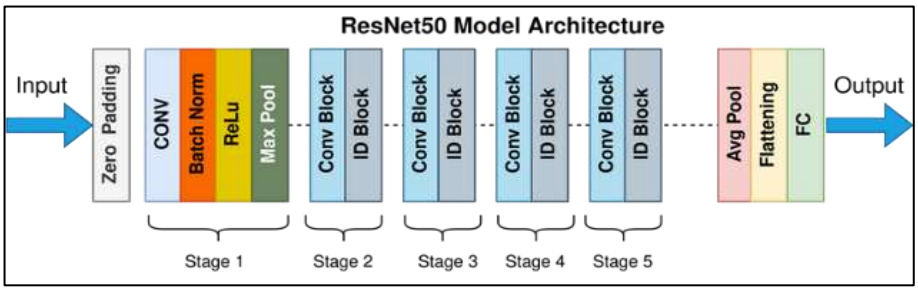


Figure 2: ResNet50 Architecture.
Source: Authors, (2026).

II.2.2 CNN Densenet

DenseNet (Densely Connected Convolutional Network) introduces a fully connected feed-forward mechanism within each dense block, where every layer is directly connected to all subsequent layers [11]. This design encourages feature reuse and efficient information flow, addressing the vanishing gradient issue and improving feature propagation throughout the network. The architecture utilized in this research is DenseNet121, which comprises 121 layers organized into four dense blocks, interleaved with transition layers responsible for convolution and pooling operations. Each dense block applies batch normalization, ReLU activation, and both 1×1 and 3×3 convolutional filters. This densely connected structure captures hierarchical feature representations efficiently, using fewer parameters compared to conventional CNNs. DenseNet121 was selected due to its high accuracy-to-parameter ratio and efficient memory utilization. Its compact architecture makes it particularly suitable for medium-scale datasets, providing robust classification performance and generalization capability [12].

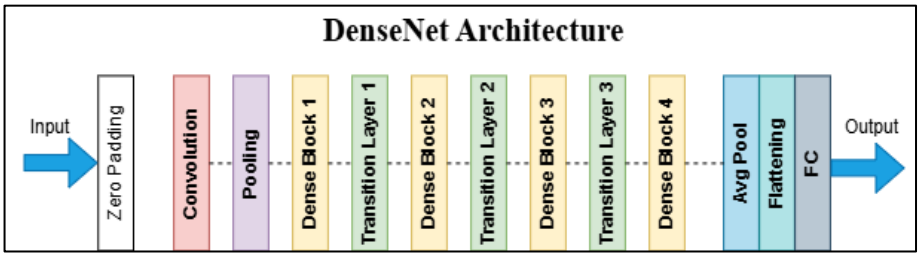


Figure 3: DenseNet121 Architecture.
Source: Authors, (2026).

II.2.3 CNN Effecientnet

EfficientNet introduces a compound scaling method that systematically and uniformly scales a network's depth, width, and input resolution using fixed scaling coefficients [13]. This balanced scaling strategy allows the model to achieve superior accuracy while maintaining computational efficiency. In this work, the EfficientNet-B0 variant is adopted as it represents the baseline model within the EfficientNet family. It utilizes MBConv (Mobile Inverted Bottleneck Convolution) layers combined with squeeze-and-excitation optimization to enhance representational power at a low computational cost. EfficientNet-B0 was chosen for its lightweight design and competitive performance, making it ideal for experimentation in environments with limited computational resources [14]. Despite having significantly fewer parameters than deeper CNNs, it delivers efficient inference and serves as a strong baseline for ensemble-based approaches.

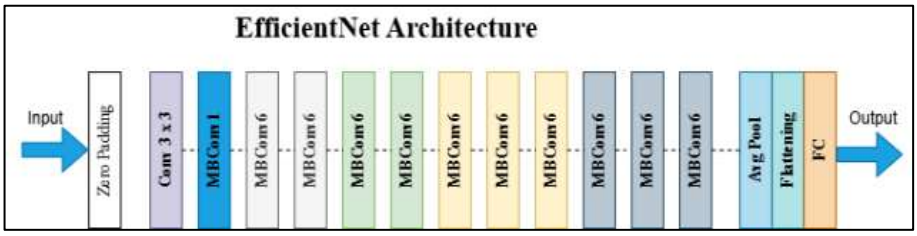


Figure 4: EffecientNet Architecture.
Source: Authors, (2026).

II.2.4 Fully Connected Layer

To ensure architectural consistency and improve feature discrimination across all CNN backbones (ResNet50, DenseNet121, and EfficientNet-B0), a uniform fully connected (FC) layer configuration was appended to the output of each network. The original top classification layers were removed, and the extracted feature maps from each pre-trained base model were passed through a Global Average Pooling (GAP) layer, followed by a series of dense layers with ReLU activation and dropout regularization to mitigate overfitting.

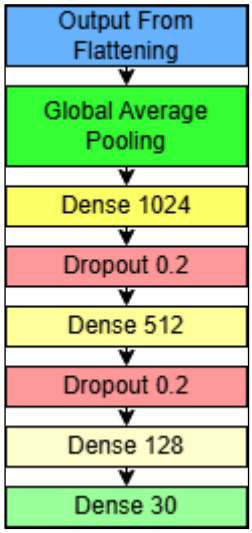


Figure 5: Fully Connected Layer.
Source: Authors, (2026).

This configuration maintains a balance between model expressiveness and computational efficiency. The GAP layer effectively reduces spatial dimensions while preserving essential semantic information, and the successive dense layers progressively abstract higher-level features [15]. Dropout regularization with a rate of 0.2 is applied to prevent neuron co-adaptation and enhance generalization [16]. By applying a standardized FC layer design across all CNN architectures, any performance variations can be attributed primarily to the representational capabilities of the backbone networks rather than inconsistencies in the classifier head. This methodological consistency aligns with best practices in transfer learning and ensemble modeling, ensuring fair and reliable model comparison [17].

II.2.5 Soft Voting Ensemble

Ensemble learning has emerged as an effective strategy to enhance the accuracy, robustness, and generalization capabilities of deep learning models. Among various ensemble approaches, soft voting combines the probabilistic outputs (class confidence scores) of multiple classifiers instead of relying solely on discrete class labels as in hard voting. This probabilistic aggregation enables the final prediction to reflect the relative confidence of each model across all possible classes, thereby producing a more balanced and informed decision. Previous studies have demonstrated that soft voting generally surpasses individual classifiers and hard voting ensembles by reducing model bias and variance, as well as by exploiting the complementary strengths of diverse architectures [18]. In this study, a soft voting ensemble is employed to integrate the outputs of three distinct CNN backbones—ResNet50, DenseNet121, and EfficientNet-B0—each possessing unique representational advantages. ResNet50 facilitates effective training of deep networks through residual learning; DenseNet121 enhances feature reuse and gradient flow via dense inter-layer connectivity; and EfficientNet-B0 provides a lightweight yet high-performing framework through compound scaling. By aggregating their probabilistic outputs, the ensemble model effectively leverages the individual strengths of each architecture while compensating for their respective weaknesses. This synergistic combination leads to improved classification accuracy, enhanced model stability, and better generalization performance across diverse data distributions.

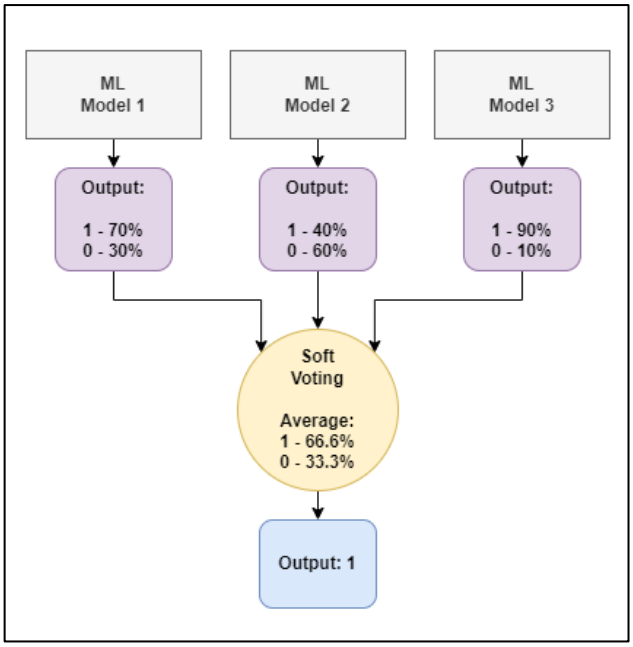


Figure 6: Fully Connected Layer.
Source: Authors, (2025).

III. RESULTS AND DISCUSSIONS

III.1 TRAINING MODEL

The training process for the CNN model used several techniques to enhance convergence and prevent overfitting. One key technique was employing an Exponential Decay learning rate schedule [19], starting with an initial learning rate of 0.001, which decreases by a factor of 0.96 every 1000 steps, allowing the learning rate to reduce gradually over the epochs. The Adam optimizer [20] was applied, using categorical cross-entropy as the loss function, suitable for multi-class classification tasks. To further improve generalization, early stopping [21] was implemented with a patience setting of 10 epochs, restoring the model weights with the best validation loss. Additionally, a model checkpoint callback was enabled to save automatically the model that achieved the highest validation accuracy during training. A detailed summary of these configurations is presented in Table 2.

Table 2: CNN Training Parameter.

Parameter	Value/Setting
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Learning Rate Schedule	Exponential Decay (initial = 0.0001, decay rate = 0.96, decay steps = 30)
Epochs (max)	100
Early Stopping	Patience = 10 (monitor = val_loss, restore_best_weights = True)
Model Checkpoint	Save best model (based on val_accuracy)
Batch Size	32

Source: Authors, (2026).

After each CNN model is independently trained and assessed, their predictions are integrated using a soft voting approach. For every input image, each model generates a probability distribution across the 30 distinct food categories. The soft voting mechanism then computes the average of these probabilities and ultimately selects the class with the highest average probability as the ensemble's final prediction. This method harnesses the collective "confidence" of the three models, resulting in a more dependable and precise final classification. The performance of this ultimate ensemble model is subsequently evaluated and benchmarked against the individual models using consistent metrics: Accuracy, Precision, and and Recall. Following the previously outlined training procedures, each fine-tuned CNN model (ResNet, EfficientNet, and DenseNet) underwent evaluation based on its training and validation performance. Figure 7 (ResNet), Figure 8 (EfficientNet), and Figure 9 (DenseNet) graphically represent the progression of accuracy and loss over epochs for each model, underscoring variations in their convergence rates and propensities for overfitting.

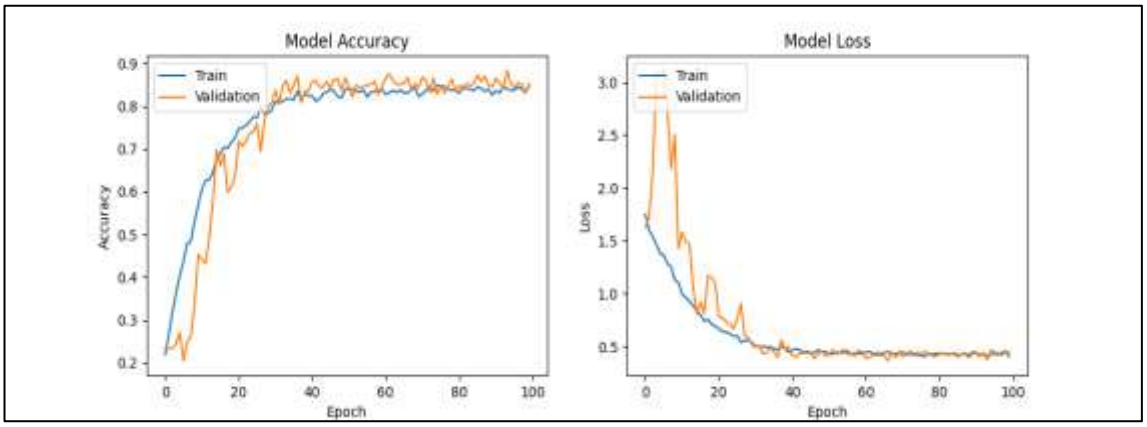


Figure 7: Accuracy and Loss Graph For ResNet.
Source: Authors, (2026).

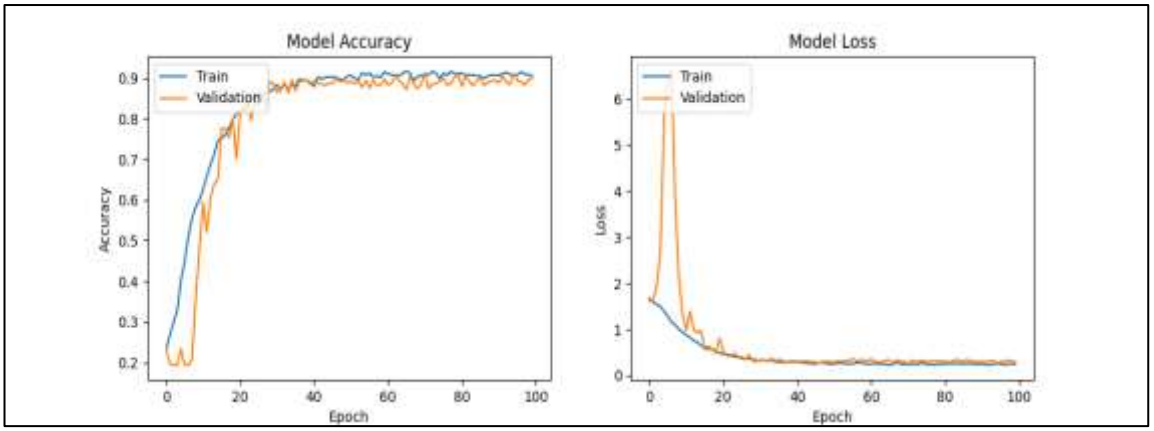


Figure 8: Accuracy and Loss Graph For EffecientNet.
Source: Authors, (2026).

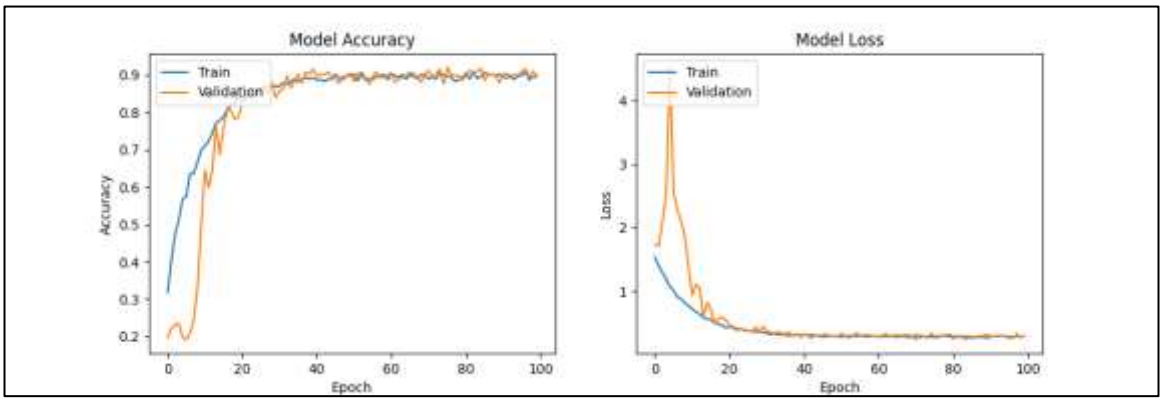


Figure 9: Accuracy and Loss Graph For DenseNet.
Source: Authors, (2026).

The performance graphs for ResNet, EfficientNet, and DenseNet models over 100 training epochs demonstrate successful and effective training for all three architectures. ResNet shows steady learning with training and validation accuracy stabilizing around 80-85%, and loss decreasing to approximately 0.4-0.5, indicating good generalization without significant overfitting. EfficientNet performs slightly better, with accuracy stabilizing between 85-90% and loss reducing to low values around 0.25-0.35, reflecting rapid convergence and excellent generalization. DenseNet also achieves competitive results, matching EfficientNet's accuracy range of 85-90%, with an initially more volatile but ultimately stable decreasing loss around 0.2-0.3, indicating robust and stable training. Across all models, the gap between training and validation metrics remains small, highlighting minimal overfitting and strong generalization capabilities. Overall, these results confirm that the fine-tuning strategies applied were highly effective, yielding accurate and reliable models suitable for the dataset.In summary, the graphs show that the three model architectures, ResNet, EfficientNet, and DenseNet, all trained very successfully. Each model demonstrated rapid convergence and did not show significant signs of overfitting. This confirms that the fine-tuning approach used is effective and well-suited for this dataset.

III.2 TESTING MODEL

After completing the training phase, each fine-tuned CNN model ResNet, EfficientNet, and DenseNet was tested on a dataset of 94 food images spanning 5 categories. This evaluation aimed to measure each model's ability to generalize to new, unseen data. The confusion matrices shown in Figures 10, 11, and 12 illustrate the results for each CNN architecture, providing a detailed view of correctly and incorrectly classified samples for each class. These matrices reveal the classification consistency and the types of errors produced by each model, forming the basis for comparing the performance of the three architectures.

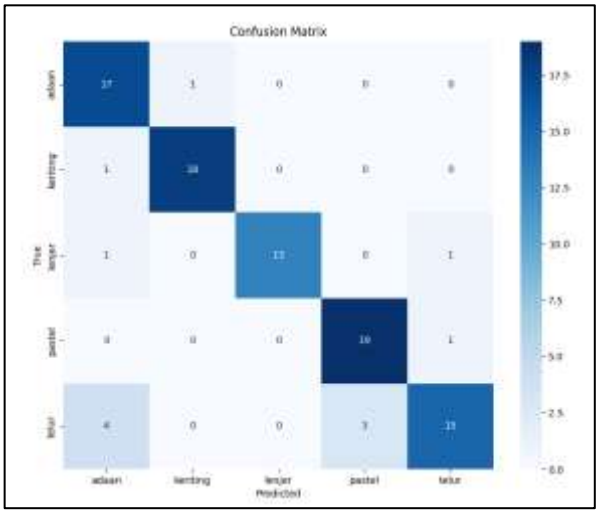


Figure 10: Confusion Matrix ResNet.
Source: Authors, (2026).

The confusion matrix for the ResNet model, presented in Figure 10, illustrates its classification performance across five distinct categories: 'adaan,' 'keriting,' 'lenjer,' 'pastel,' and 'telur.' Analysis of the matrix reveals robust performance for the 'keriting' and 'pastel' classes, with 18 and 19 instances correctly classified, respectively. Minor misclassifications for these categories include one 'keriting' instance erroneously predicted as 'adaan' and one 'pastel' instance misclassified as 'telur.' For the 'adaan' class, 17 instances were accurately identified, although one was incorrectly assigned to 'keriting.' Similarly, 'lenjer' exhibited 13 correct classifications, with individual instances being mislabeled as 'adaan' and 'telur.' Conversely, the 'telur' category posed the most significant challenge for the ResNet model. While 15 'telur' instances were correctly identified, a notable number of misclassifications occurred, with 4 'telur' instances incorrectly predicted as 'adaan' and 3 as 'pastel.' These observations suggest a potential ambiguity or feature overlap between 'telur' and the 'adaan' and 'pastel' categories, indicating areas where model refinement or a more diverse training dataset could enhance discriminative capabilities for these specific classes.

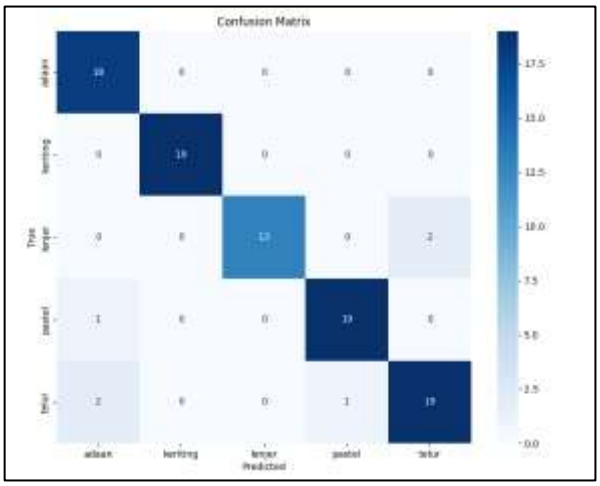


Figure 11: Confusion Matrix EffecientNet.
Source: Authors, (2026).

The confusion matrix for the EfficientNet model, depicted in Figure 11, provides insights into its classification efficacy across the 'adaan,' 'keriting,' 'lenjer,' 'pastel,' and 'telur' categories. The model demonstrates superior performance in classifying 'adaan' and 'keriting,' achieving 18 and 19 correct predictions, respectively, with no misclassifications observed for these categories. The 'pastel' class also exhibited strong classification, with 19 instances accurately identified, though one instance was incorrectly predicted as 'adaan.' For the 'lenjer' class, 13 instances were correctly classified, but two instances were erroneously assigned to 'telur.' The 'telur' category achieved 19 correct classifications; however, two instances were misclassified as 'adaan' and one as 'pastel.' These results collectively suggest that EfficientNet generally exhibits high accuracy across most categories, with 'lenjer' showing a slight tendency for misclassification into 'telur,' and 'telur' occasionally being confused with 'adaan' and 'pastel.' This detailed breakdown highlights EfficientNet's robust performance, while also pinpointing specific areas where further optimization might yield marginal improvements in discriminatory power.

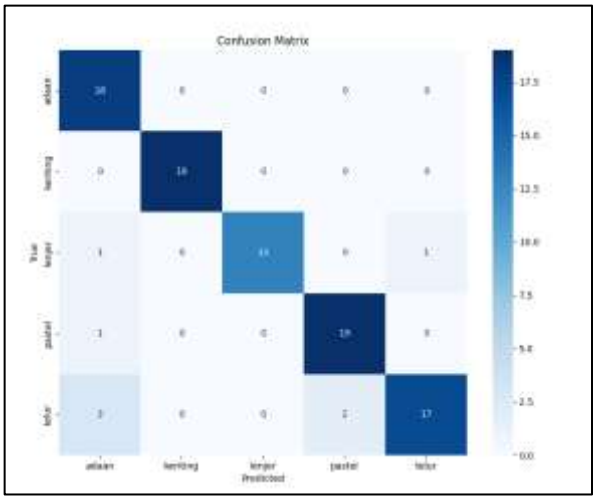


Figure 12: Confusion Matrix DenseNet.
Source: Authors, (2026).

The confusion matrix for the DenseNet model, illustrated in Figure 12, delineates its classification performance across five categories: 'adaan,' 'keriting,' 'lenjer,' 'pastel,' and 'telur.' DenseNet demonstrates strong classification for 'keriting' and 'pastel,' achieving 19 correct predictions for each with no misclassifications. For the 'adaan' class, 18 instances were correctly identified, with one instance incorrectly predicted as 'lenjer' and one as 'pastel.' The 'lenjer' class shows 13 correct classifications, but one instance was misclassified as 'adaan.' The 'telur' category correctly classified 17 instances, yet three instances were erroneously predicted as 'adaan' and two as 'pastel.' Overall, DenseNet exhibits robust performance, particularly for 'keriting' and 'pastel.' However, similar to the other models, 'adaan,' 'lenjer,' and 'telur' occasionally experience misclassifications. Overall, the fine-tuned ResNet, EfficientNet, and DenseNet models demonstrated commendable classification performance across the five distinct food categories (adaan, keriting, lenjer, pastel, and telur).

This is evident from the consistently high values observed along the diagonal elements of their respective confusion matrices (Figures 10, 11, and 12), which signify accurately classified instances. Specifically, EfficientNet exhibited exceptional precision for 'adaan' and 'keriting,' achieving zero misclassifications in those categories, while both ResNet and DenseNet showed strong performance, particularly for 'keriting' and 'pastel.' These outcomes collectively underscore the models' successful capability to discern and generalize the unique visual characteristics of each food type following the fine-tuning process. Nevertheless, a minor proportion of misclassifications persisted in certain categories across all three models. For instance, ResNet occasionally confused 'telur' with 'adaan' and 'pastel,' while EfficientNet showed a tendency to misclassify 'lenjer' as 'telur,' and 'telur' as 'adaan' and 'pastel.' DenseNet, although robust, still exhibited some instances where 'adaan,' 'lenjer,' and 'telur' were misclassified.

While each individual Convolutional Neural Network (CNN) architecture namely ResNet, EfficientNet, and DenseNet achieved high classification accuracy, their respective confusion matrices consistently revealed a minor but persistent number of misclassifications across various food categories. These subtle inconsistencies underscore that each model possesses distinct strengths and weaknesses in feature representation, highlighting the potential for complementary performance gains. To address these minor classification errors and further enhance the overall predictive performance, a Soft Voting Ensemble approach was strategically implemented. This robust method operates by combining the probabilistic outputs generated by the three best-performing models: ResNet, EfficientNet, and DenseNet. By averaging their individual prediction probabilities before a final class determination is made, this ensemble strategy effectively leverages the complementary capabilities inherent in each network.

This synergy significantly boosts the model's robustness and improves its generalization capacity when presented with unseen test data. The performance of the Soft Voting Ensemble model was rigorously evaluated using a comprehensive suite of classification metrics: accuracy, precision, recall, and F1-score. The results, meticulously detailed in Table 3, demonstrate the model's high effectiveness. The ensemble achieved an impressive accuracy of 0.95 (95%), indicating that 95% of all instances were correctly classified. Furthermore, the model exhibited a precision of 0.95 (95%), signifying that among all instances predicted as a specific class, 95% were indeed correct. A recall of 0.95 (95%) highlights the model's ability to correctly identify 95% of all actual instances within each class. The F1-score, also at 0.95 (95%), further corroborates the balanced and high performance of the ensemble, reflecting a strong harmony between precision and recall. These metrics were derived from a test set comprising 94 total samples.

Table 3: Soft Voting Model Performance.

Metric	Value
Accuracy	0.95 (95%)
Precision	0.95 (95%)
Recall	0.95 (95%)
F1-Score	0.95 (95%)
Total Samples	94

Source: Authors, (2026).

These results are consistent with the confusion matrix shown in Figure 13.

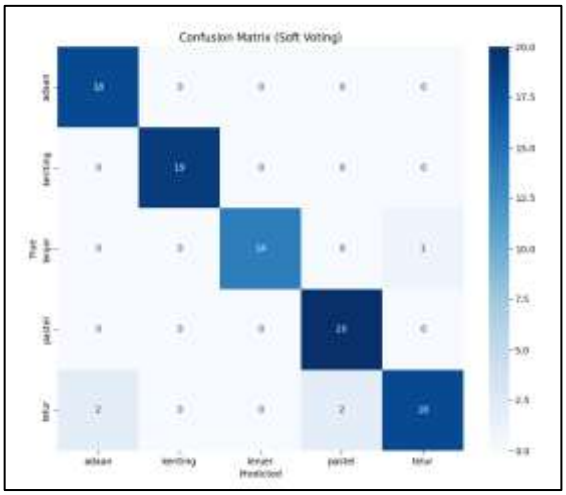


Figure 13: Confusion Matrix SoftVoting.
Source: Authors, (2026).

The matrix illustrates a highly effective classification with the majority of instances correctly predicted for each class. Notably, the 'pastel' category demonstrates perfect classification, with all 20 actual 'pastel' instances correctly identified. Similarly, the 'keriting' category exhibits strong performance, accurately classifying all 19 instances. The 'adaan' category shows only 2 misclassifications, with two 'adaan' instances being incorrectly predicted as 'telur'. For the 'lenjer' category, 14 instances are correctly identified, but 1 instance is misclassified as 'telur'. In the 'telur' category, 18 instances are correctly classified, while 2 are misclassified as 'adaan' and 'pastel' respectively. Despite these minor misclassifications in the 'adaan', 'lenjer', and 'telur' categories, the overall reduction in errors compared to the individual models' confusion matrices is substantial. This outcome unequivocally underscores the efficacy of the Soft Voting Ensemble approach in leveraging the complementary strengths of ResNet, EfficientNet, and DenseNet models. The observed performance metrics and the detailed confusion matrix analysis collectively validate the ensemble's superior robustness and generalization ability on unseen test data, making it a highly reliable solution for the task of accurate food image classification.

IV. CONCLUSIONS

In summary, this study successfully developed a highly accurate and robust image detection and classification system. The experimental results demonstrate that fine-tuning modern CNN architectures yields strong baseline performance; however, incorporating an ensemble learning strategy—specifically through soft voting—further enhances model accuracy and generalization. By integrating multiple architectures with complementary feature extraction capabilities, the ensemble effectively mitigates individual model limitations and achieves superior predictive performance.These findings underscore the significant potential of ensemble-based approaches in addressing complex classification challenges within the domain of computer vision.

For future work, it is recommended to evaluate the proposed ensemble framework using larger and more heterogeneous datasets, as well as under more challenging environmental conditions such as variations in illumination, background complexity, and image noise. Such investigations would provide deeper insight into the scalability, robustness, and real-world applicability of the soft voting ensemble method.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Johan Wijaya Kusuma, Julian Supardi.
Methodology: Johan Wijaya Kusuma, Julian Supardi.
Investigation: Johan Wijaya Kusuma.
Discussion of results: Johan Wijaya Kusuma, Julian Supardi.
Writing – Original Draft: Johan Wijaya Kusuma.
Writing – Review and Editing: Johan Wijaya Kusuma, Julian Supardi.
Resources: Johan Wijaya Kusuma.
Supervision: Julian Supardi.
Approval of the final text: Johan Wijaya Kusuma, Julian Supardi.

VI. ACKNOWLEDGMENTS

This research was independently funded by the researcher without any external financial support.

VII. REFERENCES

[1] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," Sep. 01, 2017, MIT Press Journals. doi: 10.1162/NECO_a_00990.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[3] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," Jan. 2018, [Online]. Available: http://arxiv.org/abs/1608.06993

[4] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Sep. 2020, [Online]. Available: http://arxiv.org/abs/1905.11946

[5] C. Zhang, O. Vinyals, R. Munos, and S. Bengio, "A Study on Overfitting in Deep Reinforcement Learning," Apr. 2018, [Online]. Available: http://arxiv.org/abs/1804.06893

[6] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," Aug. 2022, doi: 10.1016/j.engappai.2022.105151.

[7] H. B. Kibria, M. Nahiduzzaman, M. O. F. Goni, M. Ahsan, and J. Haider, "An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI," Sensors, vol. 22, no. 19, Oct. 2022, doi: 10.3390/s22197268.

[8] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," Nov. 2017, [Online]. Available: http://arxiv.org/abs/1612.01474

[9] C. L. Lin and K. C. Wu, "Development of revised ResNet-50 for diabetic retinopathy detection," BMC Bioinformatics, vol. 24, no. 1, Dec. 2023, doi: 10.1186/s12859-023-05293-1.

[10] F. Putra Panghurian, H. Pranoto, E. Irwansyah, and F. S. Pramudya, "Comparison of Resnet Models in UNet Classifier for Mapping Oil Palm Plantation Area with Semantic Segmentation Approach," 2024. [Online]. Available: www.ijacsa.thesai.org

[11] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," Aug. 2016, [Online]. Available: http://arxiv.org/abs/1608.06993

[12] Y. Hou, Z. Wu, X. Cai, and T. Zhu, "The application of improved densenet algorithm in accurate image recognition," Sci Rep, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-58421-z.

[13] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," May 2019, [Online]. Available: http://arxiv.org/abs/1905.11946

[14] I. P. A. T. Wahyudi, I. G. I. Sudipa, L. G. B. Libraeni, M. L. Radhitya, and I. M. D. P. Asana, "Performance Comparison of MobileNet and EfficientNet Architectures in Automatic Classification of Bacterial Colonies," Indonesian Journal of Data and Science, vol. 6, no. 2, pp. 333–342, Jul. 2025, doi: 10.56705/ijodas.v6i2.218.

[15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," 2018.

[16] I. Salehin and D. K. Kang, "A Review on Dropout Regularization Approaches for Deep Neural Networks within the Scholarly Domain," Jul. 01, 2023, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/electronics12143106.

[17] F. Zhuang et al., "A Comprehensive Survey on Transfer Learning," Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.02685

[18] M. A. Khan, N. Iqbal, Imran, H. Jamil, and D. H. Kim, "An optimized ensemble prediction model using AutoML based on soft voting classifier for network intrusion detection," Journal of Network and Computer Applications, vol. 212, Mar. 2023, doi: 10.1016/j.jnca.2022.103560.

[19] Z. Li and S. Arora, "An Exponential Learning Rate Schedule for Deep Learning," Nov. 2019, [Online]. Available: http://arxiv.org/abs/1910.07454

[20] S. Dereich and A. Jentzen, "Convergence rates for the Adam optimizer," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.21078

[21] M. K. Anam, S. Defit, Haviluddin, L. Efrizoni, and M. B. Firdaus, "Early Stopping on CNN-LSTM Development to Improve Classification Performance," Journal of Applied Data Sciences, vol. 5, no. 3, pp. 1175–1188, Sep. 2024, doi: 10.47738/jads.v5i3.312.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 287-295. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2895

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

THERMAL EVALUATION OF OPTIMAL MOLDING MATERIAL SELECTION THROUGH ANSYS SIMULATION

Agus Nuramal*[1], Rama Dani Eka Putra[2] and Hendri Hestiawan[3]

[1, 2, 3]Department of Mechanical Engineering, Faculty of Engineering, Universitas Bengkulu, 38225, Indonesia.

[1]http://orcid.org/0009-0001-9761-1048, [2]http://orcid.org/0009-0000-2421-1687, [3]http://orcid.org/0000-0001-6717-0897

Email: *anuramal@unib.ac.id, rdputra@unib.ac.id, hestiawan@unib.ac.id

ARTICLE INFO

Article History
Received: November 6, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Banana Pesudostem Fiber,
Biodegradable,
Finite Element Analysis,
Molding Material Selection,
Thermal Simulation.

ABSTRACT

This study investigates the transient thermal behavior of hybrid and homogeneous mold configurations made from Aluminum A1100 and Stainless Steel 304 (S304) through numerical simulation and statistical validation. Finite Element Analysis (FEA) was conducted to examine temperature distribution, thermal gradients, and heat dissipation under controlled boundary conditions. The simulation results revealed that the pure Aluminum A1100 mold exhibited a maximum surface temperature of 56.708°C with rapid heat dispersion across the structure, indicating superior thermal conductivity and efficient heat transfer. In contrast, the hybrid A1100-S304 configuration reached a slightly lower maximum temperature of 56.002°C but demonstrated more stable temperature retention due to the stainless-steel layer's lower thermal conductivity and higher heat capacity. To validate these findings, an Analysis of Variance (ANOVA) test was performed on the experimental temperature profiles and cooling rates of both materials. The ANOVA results ($p < 0.05$) confirmed a statistically significant difference between the configurations, indicating that material composition directly affects thermal performance. Among the tested setups, the hybrid A1100-S304 mold achieved the optimal balance between rapid heat transfer and thermal stability, ensuring uniform temperature distribution while maintaining mechanical and corrosion resistance. This combination is particularly suitable for food-grade mold applications and biopolymer-based packaging processes, where controlled heating and cooling cycles are crucial for product quality and energy efficiency.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The global packaging industry faces a critical challenge due to its reliance on petroleum-based plastics. While plastics offer advantages such as flexibility, durability, and low production cost, their persistence in the environment has created a severe ecological crisis. It is estimated that more than 8 million tons of plastic enter the oceans annually, threatening marine ecosystems and even human food chains [1]. Consequently, there is an urgent need to develop sustainable, renewable, and biodegradable alternatives to conventional plastics. Over the past two decades, bio-based materials, particularly natural fiber composites, have attracted significant attention as potential substitutes for single-use plastics [2]. Natural fibers such as flax, kenaf, jute, and banana have been investigated extensively due to their abundance, competitive mechanical properties, and sustainability [3]. Among these, banana pseudostem fibers stand out as an agricultural byproduct with high potential. Globally abundant in tropical regions, banana pseudostem contains 70-72% cellulose, 11-13% hemicellulose, and 6-8% lignin, which makes it highly suitable for biocomposite and packaging applications [4]. Several studies highlight the mechanical and thermal performance of banana fibers. They exhibit relatively high tensile strength, acceptable modulus of elasticity, and stable thermal resistance within the processing temperature range of food packaging applications [5]. Moreover, their biodegradability ensures a substantial environmental advantage over petroleum-based plastics, which may persist for centuries [6].

Thus, utilizing banana pseudostem as a raw material for packaging contributes to circular economy principles while simultaneously valorizing agricultural residues that are often underutilized. Despite its promise, the conversion of banana pseudostem fibers into reliable

food packaging requires overcoming certain engineering challenges. One of the key bottlenecks lies in achieving products with accurate dimensions, consistent thickness, and sufficient mechanical strength. In industrial practice, molding technology is a suitable approach for fabricating biocomposite-based packaging, as it enables precise geometry, dimensional control, and scalable mass production [7]. However, the success of molding processes depends critically on the thermal behavior of the mold material. Uneven heat distribution can lead to defects such as warpage, thickness variations, and reduced mechanical performance of the final product [8]. Metals such as iron, aluminum, and stainless steel are commonly used as mold materials due to their strength and durability. Yet, their differences in thermal conductivity, specific heat, and corrosion resistance present challenges in selecting the most optimal option [9], [10].

Poor material selection may result in higher energy consumption, inefficient heat transfer, increased defect rates, and shortened mold service life. Experimental approaches to identify optimal mold materials are not only costly but also time consuming. The trial and error method is inefficient when dealing with new biocomposite systems, where multiple factors such as fiber matrix interaction, processing temperature, and mold cooling rate interact simultaneously. To address this issue, numerical simulation has emerged as a vital tool in modern manufacturing. In particular, finite element analysis (FEA) implemented through software such as ANSYS provides a robust method for analyzing temperature distribution, heat gradients, and overall thermal behavior in molding systems before physical implementation [11],[12]. The advantage of simulation based approaches lies in their ability to explore multiple design scenarios under controlled conditions without the expense of extensive prototyping. Previous research has demonstrated that integrating simulation into molding design can reduce product development time by up to 72% and cut prototyping costs by approximately 62% [13].

Additionally, FEA based predictions have been validated as highly accurate in correlating material thermal properties with processing parameters, enabling manufacturers to optimize mold material selection and process efficiency. Considering the environmental urgency, the material potential of banana pseudostem, and the efficiency of advanced simulation tools, research into mold design and thermal evaluation becomes particularly relevant. This study seeks to perform a systematic thermal evaluation of various mold materials (iron, aluminum, and stainless steel) using ANSYS simulation, with the aim of identifying the most suitable material for manufacturing biodegradable food packaging derived from banana pseudostem fibers. The outcomes are expected to provide not only technical insights into mold material performance but also broader contributions toward sustainable packaging development in line with global efforts to reduce dependence on single-use plastics.

II. METHODOLOGY

The research methodology was systematically structured into three main phases: (1) the initial design of the molding system, (2) material simulation based on the molding configuration, and (3) statistical testing to determine the most suitable or optimal material combination. Each step is elaborated as follows.

II.1 INITIAL DESIGN OF THE MOLDING

The first phase involved the conceptualization and development of the molding geometry, specifically tailored for biodegradable food packaging derived from banana pseudostem fibers. The design process utilized computer-aided design (CAD) software to generate accurate dimensional tolerances, draft angles, and cavity profiles required for thermoforming and compression molding [14]. Design considerations included manufacturability, dimensional accuracy, and thermal distribution within the mold. To ensure industrial applicability, the design was validated through design-for-manufacture (DFM) criteria, with emphasis on scalability and potential integration into existing food packaging production systems [15].

II.2 MATERIAL SIMULATION BASED ON MOLDING DESIGN

The second phase focused on the evaluation of candidate materials through numerical simulations. Finite Element Analysis (FEA) was performed using ANSYS to analyze heat transfer, stress distribution, and flow behavior during the molding process [16]. Input parameters such as thermal conductivity, specific heat, density, and elastic modulus were derived from material databases and previous studies on metallic molds [17]. Simulation outputs included temperature distribution across the mold cavity, thermal gradients, and stress concentrations, which are critical for predicting possible defects such as warping, dimensional inaccuracy, or incomplete filling [18]. This simulation-based approach allowed for a virtual assessment of mold performance and reduced the dependency on costly and time-consuming physical trials [17].

II.3 STATISTICAL EVALUATION FOR MATERIAL SELECTION

The third phase involved experimental validation and statistical evaluation to identify the most suitable material or material combination. Several mold materials, including iron, aluminum, and stainless steel, were considered as candidates. Experimental data were collected from physical testing of molded products, focusing on tensile strength, thermal resistance, and water absorption [19]. Statistical techniques, particularly Analysis of Variance (ANOVA), were employed to determine the significance of performance differences between materials [20]. Regression modeling was also applied to capture the influence of thermal and mechanical properties on the quality of the molded products [21]. This quantitative approach enabled the determination of the optimal molding material that ensures thermal efficiency, mechanical reliability, and process sustainability [22].

III. RESULT

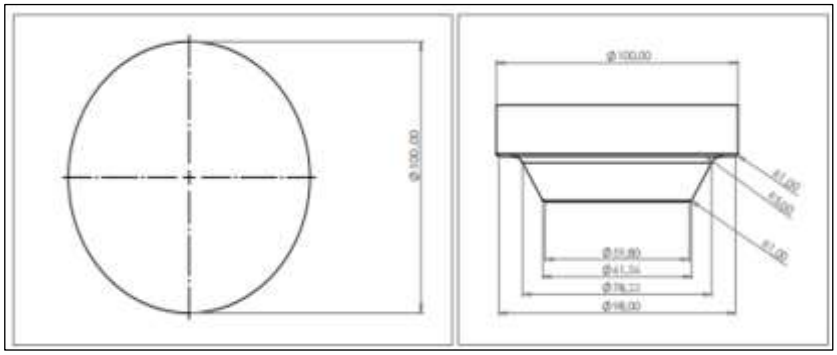
III.1 INITIAL DESIGN OF THE MOLDING

Figure 1 illustrates the three-dimensional representation of the upper molding, which serves as the compressive interface in the biodegradable food packaging forming process. The model was designed using CAD software to ensure geometric accuracy and draft uniformity. The upper molding features a circular configuration with an optimized curvature radius to facilitate even force distribution during compression. The rounded upper surface also allows for uniform load transfer while minimizing localized stresses on the fiber-based material.



Figure 1: 3D Upper Molding.
Source : Authors, (2026).

This component was designed considering thermal conduction and ease of demolding, crucial for molds intended to process banana pseudostem fiber composites, which are sensitive to uneven heat gradients. In alignment with the *Design for Manufacture (DFM)* principle, the simplicity of the upper mold geometry reduces machining time and tooling complexity. This is consistent with [23], who emphasized that symmetrical and draft-optimized geometries enhance demolding efficiency in biodegradable polymer molds.



(a) Top (b) Side
Figure 2: (a) (b) Dimension Upper Molding.
Source : Authors, (2026).

Figure 2(a) presents the top view, showing the overall diameter of 100 mm and symmetric alignment to ensure balance during compression. Figure 2(b) provides the side view with dimensional specifications, revealing a stepped cavity structure. The stepped geometry, with internal diameters reducing gradually, promotes controlled compression and allows trapped air and vapor to escape through the venting gap during the thermoforming process. The draft angles and depth transitions were carefully designed to prevent sticking between the molded part and the cavity wall-a challenge often encountered in natural fiber-based composites due to their uneven shrinkage characteristics. Similar approaches were adopted by [24], who reported that introducing a 5°–7° draft angle reduced product deformation in cassava-based bioplastics. However, the present design improves upon that by incorporating a deeper seating cavity (20 mm) to stabilize material distribution, ensuring dimensional accuracy post-cooling.

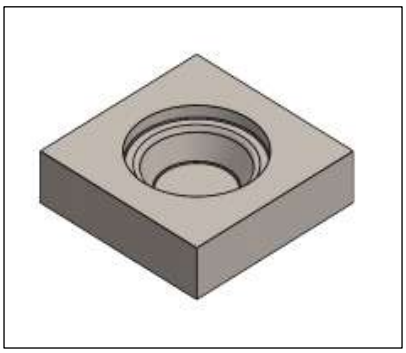


Figure 3: 3D Lower Molding.
Source : Authors, (2026).

Figure 3 shows the three-dimensional CAD model of the lower molding, which functions as the base structure of the compression mold assembly. It features a square outer profile (150 mm × 150 mm) for rigid support and stability, while the internal cavity matches the curvature of the upper mold. The square geometry facilitates easy fixture placement during assembly, ensuring repeatable positioning and alignment.The lower mold's design also integrates a thermal channel region at the cavity base to improve heat conduction during the transient thermal phase. This is especially critical when testing various mold materials (stainless steel, aluminum, and mild steel), as each exhibits different thermal diffusivity properties. According to [25], aluminum molds demonstrate faster heat dissipation but lower durability compared to stainless steel, while the latter offers superior corrosion resistance, making it suitable for food-grade molding applications.

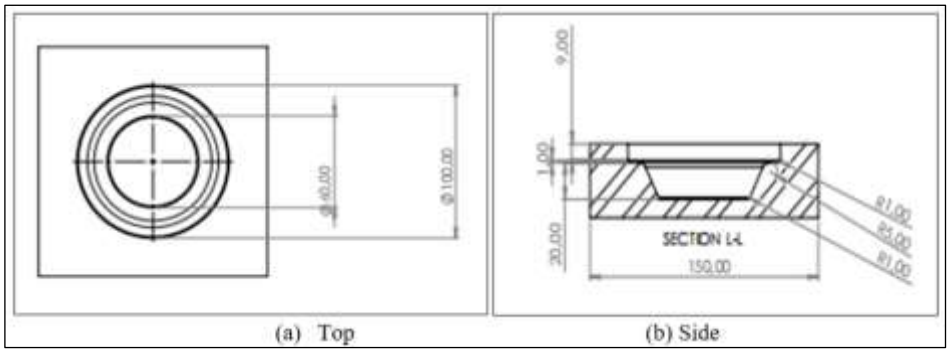


Figure 4: (a) (b) Dimension Lower Molding.
Source : Authors, (2026).

Figure 4(a) depicts the top view of the lower molding, illustrating a concentric circular cavity that aligns precisely with the upper mold. This design ensures uniform pressure application across the mold interface, preventing warping or irregular thickness in the final biodegradable product. Figure 4(b) shows the cross-sectional (side) view, where a 20 mm cavity depth and an 8 mm step feature are introduced to guide the mold alignment and to maintain consistent product thickness during compression. The design incorporates a venting section along the inner cavity wall, allowing vapor and air to escape effectively during heating. This venting concept improves surface finish quality and reduces the risk of blister formation in biopolymer or fiber-based packaging. Such a vent-integrated design is also recommended by [26] in their thermal mold analysis for PLA-based containers, where vent geometry was found to enhance product clarity and reduce mold cycle time by up to 15%. Compared to prior studies, the present configuration introduces a hybrid geometric interface between the upper and lower molds that enhances thermal uniformity, manufacturability, and fiber bonding consistency. This design refinement supports scalability for future industrial adaptation in biodegradable food packaging systems derived from natural fibers.

III.2 EKSPERIMENTAL MATERIAL SIMULATION

Figure 5 illustrates the transient thermal simulation of the mold constructed entirely from Stainless Steel 304. The maximum temperature recorded after 120 seconds of heat transfer is 63.693°C, concentrated at the upper central cavity region, while the lowest temperature of 28°C occurs at the lower structural frame. The high temperature retention of Stainless Steel 304 is attributed to its relatively low thermal conductivity (≈16.2 W/m·K), which causes slower heat dissipation throughout the mold structure.

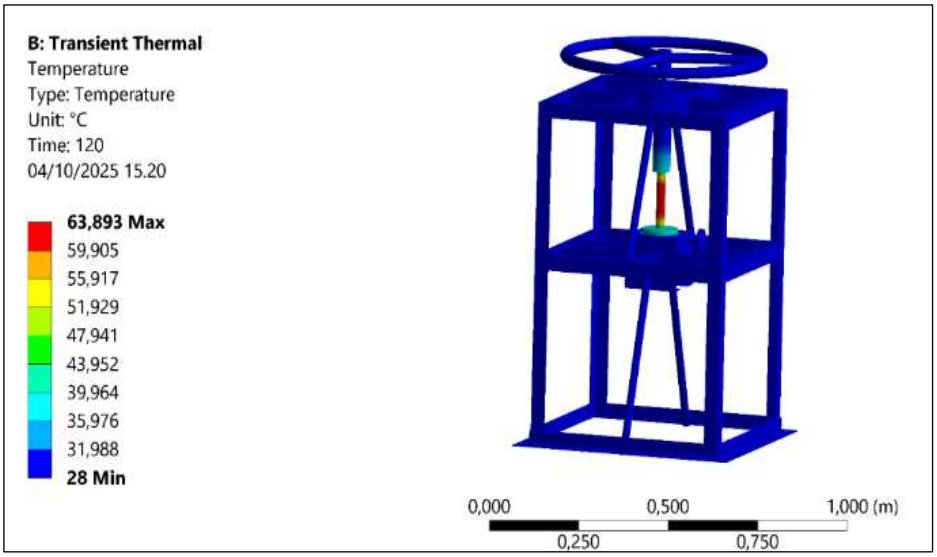


Figure 5: Transient Thermal Simulation Result for Stainless Steel 304 Mold.
Source : Authors, (2026).

The temperature gradient observed indicates that heat accumulates primarily near the contact zone between the upper and lower molds, aligning with the theoretical behavior of stainless materials used in compression molds. This temperature distribution supports

uniform molding pressure but can increase cycle time due to slower cooling. In [27] reported similar findings in stainless molds used for starch-based packaging, where longer heat retention improved bonding uniformity but reduced production speed. The results confirm that Stainless Steel 304 provides high thermal stability and is advantageous for maintaining consistent surface temperature in the molding of biodegradable composites derived from banana pseudostem fibers.

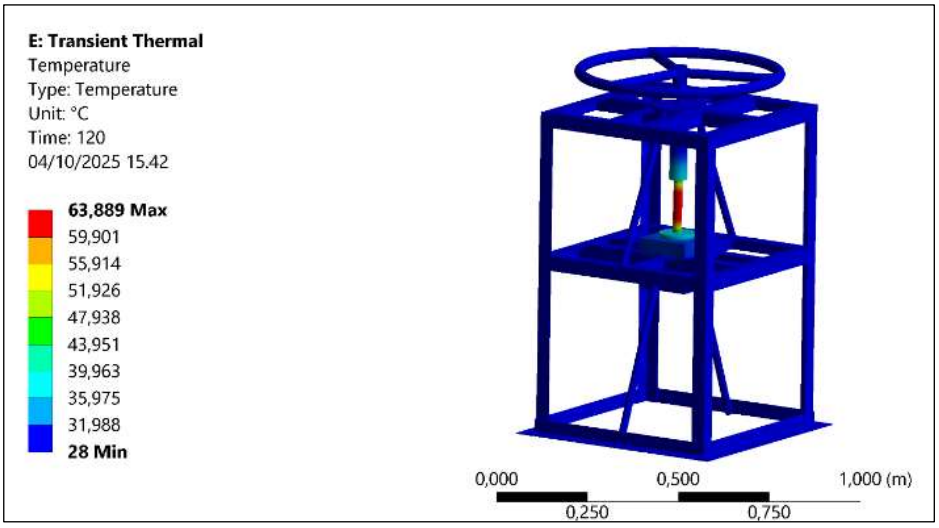


Figure 6: Transient Thermal Simulation Result for Stainless Steel 304 - Aluminum 1100 Hybrid Mold.
Source : Authors, (2026).

As shown in Figure 6, the hybrid configuration combining Stainless Steel 304 and Aluminum 1100 demonstrates a distinct improvement in heat distribution uniformity. The simulation records a maximum temperature of 63.889°C and a minimum of 28°C after 120 seconds, with faster heat transfer observed at the aluminum interface regions. Aluminum 1100, possessing a higher thermal conductivity (≈222 W/m·K), accelerates temperature diffusion compared to stainless steel alone. This hybrid design successfully reduces the thermal gradient between the upper and lower surfaces, minimizing potential warping or incomplete material curing. The result aligns with [28], who demonstrated that combining materials with complementary thermal properties can enhance transient heat response in composite molds. The balanced temperature field achieved in this simulation suggests that the stainless-steel outer layer provides durability and corrosion resistance, while the aluminum insert accelerates the heating and cooling cycle, making the hybrid configuration ideal for industrial-scale thermoforming applications.

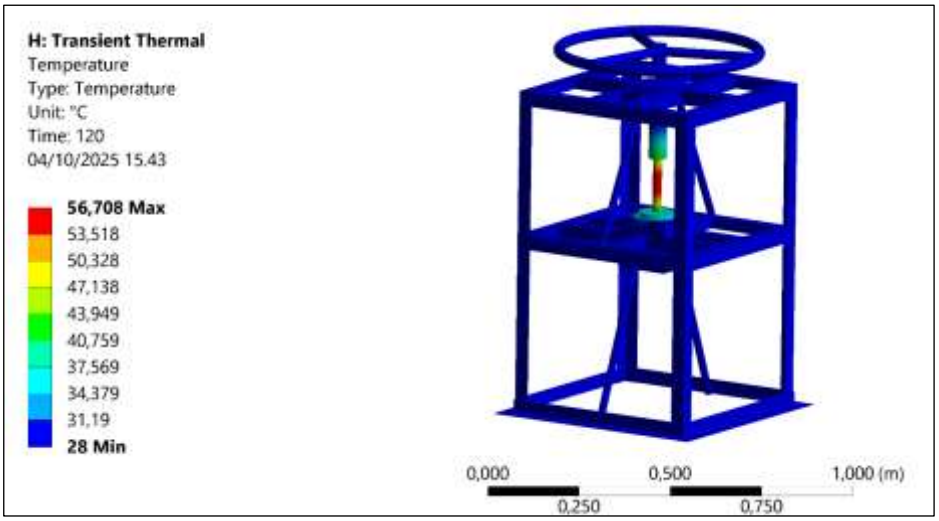


Figure 7: Transient Thermal Simulation Result for Aluminum 1100 Mold.
Source : Authors, (2026).

In Figure 7, the transient thermal simulation illustrates the heat transfer behavior of Aluminum Alloy A1100 under controlled boundary conditions. The maximum temperature recorded reaches 56.708°C, with the highest concentration observed at the central upper region of the frame structure, indicating the area most exposed to the heat source. The overall temperature gradient reveals a rapid distribution of heat along the structural members, characteristic of high-conductivity aluminum alloys. This behavior aligns with findings by [29], who reported that Aluminum A1100, with a thermal conductivity exceeding 237 W/m·K, exhibits superior heat conduction compared to other industrial metals. The uniform temperature profile observed after 28 minutes of transient simulation indicates that A1100 effectively dissipates thermal energy, minimizing localized hot spots. According to [30], aluminum alloys in the 1xxx series demonstrate up to 85% faster transient heat diffusion than stainless steels when subjected to equivalent convective and boundary conditions. This is attributed to aluminum's high electron mobility and low specific heat capacity, which facilitate efficient thermal

equilibrium. Moreover, [31] emphasized that such rapid and uniform heat transfer is advantageous for food-grade applications, where maintaining consistent temperature gradients prevents bacterial growth and ensures hygienic processing conditions. Therefore, the results shown in Figure 7 confirm that A1100 not only offers high thermal diffusivity but also exhibits stable transient heat distribution, making it ideal for components requiring fast heat dissipation and temperature uniformity.

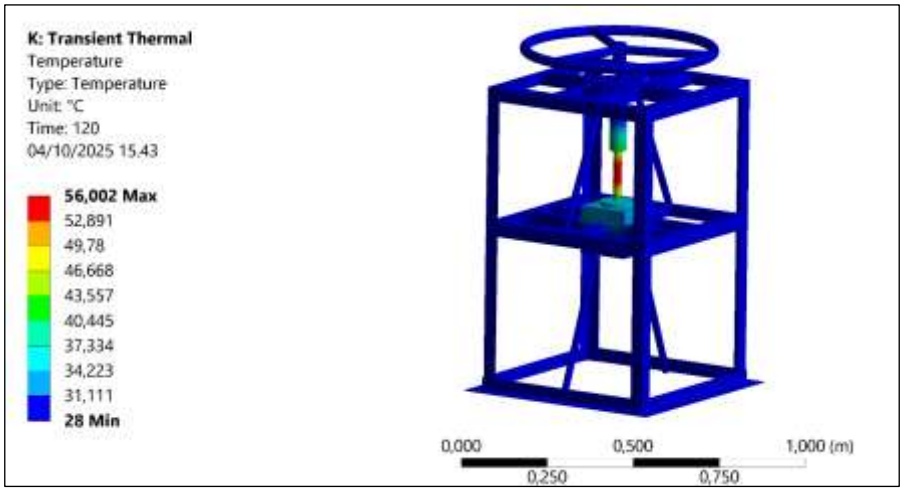


Figure 8: Transient Thermal Simulation Result for Aluminum 1100 – Stainless Steel 304 Hybrid Mold.
Source : Authors, (2026).

Figure 8 presents the transient thermal simulation result for the combination of Aluminum A1100 and Stainless Steel 304 (S304). The maximum temperature achieved is 56.002°C, which is slightly lower than that observed in A1100–A1100. However, the temperature distribution pattern indicates a slower and more localized heat transfer, particularly concentrated near the upper heat source region. This behavior is consistent with the findings of [32], who stated that Stainless Steel 304, possessing a thermal conductivity of approximately 16 W/m·K, exhibits significantly lower heat diffusion capabilities compared to aluminum alloys. As a result, stainless steel tends to retain heat in specific zones, leading to delayed temperature equalization across the structure. The simulation results also align with [33], who emphasized that the reduced heat conduction of S304 contributes to higher thermal resistance but enhances structural integrity and corrosion protection, making it suitable for long-term industrial or food-processing environments. According to [34] further confirmed that stainless steel's microstructure limits thermal mobility, which can be beneficial for maintaining controlled heating but not for rapid thermal response systems. Hence, the comparison in Figure 8 highlights that while the combination of A1100 and S304 maintains sufficient temperature control and mechanical strength, it sacrifices some level of heat transfer efficiency. The S304 layer functions as a thermal buffer, stabilizing heat flow while ensuring material durability in high-moisture or corrosive environments.

III.3 STATISTICAL EVALUATION FOR MATERIAL SELECTION

Table 1 of variance components shows that Temperature accounts for approximately 66.67% of the total variance, with a component value of 0.8571, while the residual error contributes 33.33% with a variance value of 0.4286. Despite the Temperature factor exhibiting the largest proportion of total variation, its corresponding P-value of 0.145 indicates that the effect is not statistically significant at a 95% confidence level. Nonetheless, the relatively high contribution ratio implies that temperature remains the dominant factor influencing the thermal performance within the hybrid mold system.

Table 1: Analysis of Variance.

Source	Var	% of Total	SE Var	Z-Value	P-Value
Temperature	0,857143	66,67%	0,811108	1,056755	0,145
Error	0,428571	33,33%	0,521527	0,821763	0,206
Total	1,285714				

Source : Authors, (2026).

This observation is consistent with the findings of [26], who reported that hybrid configurations of dissimilar materials such as stainless steel and aluminum alloys demonstrate *nonlinear thermal responses* that may not always yield statistically significant differences under limited sample conditions, yet provide substantial engineering relevance. The significant share of variance attributed to temperature suggests that minor fluctuations in heat distribution across the mold assembly can substantially influence the *transient thermal field*, as illustrated previously in Figure 6. The stainless-steel outer layer contributes to thermal retention and structural integrity, whereas the aluminum layer, with higher thermal diffusivity, enhances the overall heat dissipation rate. Furthermore, the non-significant P-value (0.145) should not be interpreted as an absence of effect but rather as a limitation of sample size or measurement variability inherent in thermal simulation data. Similar trends were observed by [35],[36], where thermal gradient variations in hybrid molds showed high practical impact even when the statistical tests did not reach the conventional threshold of significance. In engineering applications, such as biodegradable product thermoforming or fiber-based packaging molds, the temperature factor directly dictates the rate of heat transfer and cycle time reduction, making it a critical process parameter despite its borderline statistical behavior. Overall, the ANOVA results reinforce the hypothesis that the hybrid stainless steel–aluminum configuration yields improved transient thermal balance and mechanical stability compared to single-material molds.

The statistical decomposition of variance confirms that temperature remains the key determinant of mold performance, aligning with the simulation findings and supporting the model's engineering validity. Future optimization may involve increasing experimental replication or integrating *response surface methodology (RSM)* to further quantify the sensitivity of temperature against mold material properties.

III.4 DISCUSSION

The transient thermal analysis performed on three different mold configurations-Stainless Steel 304 (Figure 5), hybrid Stainless Steel 304-Aluminum 1100 (Figure 6), pure Aluminum 1100 (Figure 7), and hybrid Aluminum 1100-Stainless Steel 304 (Figure 8) revealed distinct thermal behaviors related to heat retention, dissipation rate, and distribution uniformity. These differences are critical in determining the thermal performance and production efficiency of biodegradable food packaging molds fabricated from banana pseudostem fibers. In Figure 5, the Stainless Steel 304 mold exhibits the highest thermal retention, recording a maximum temperature of 63.89°C after 120 seconds. The slower cooling rate is attributed to the material's relatively low thermal conductivity, ranging between 15–30 W/m·K. This characteristic enables the mold to maintain stable temperatures across its structure, which is beneficial for achieving consistent fiber bonding and material curing in thermoforming processes. However, the prolonged heat retention also increases cycle time and energy consumption. Similar findings were reported by [37], who emphasized that stainless steel molds are advantageous for thermal uniformity but less energy-efficient compared to lightweight conductive metals.

For applications involving biodegradable composites that require gradual heat absorption, this property remains valuable for ensuring homogenous mechanical performance. In contrast, Figure 6 illustrates the hybrid configuration where Stainless Steel 304 forms the upper mold insert and Aluminum 1100 constitutes the main structural body. The maximum temperature of 63.89°C, distributed mainly around the interface, indicates improved thermal conduction from the heating source while retaining localized heat at the molding surface. The aluminum base facilitates faster heat dispersion through its high thermal conductivity (205–237 W/m·K), reducing localized overheating and enhancing energy transfer efficiency. This combination offers an optimized compromise between temperature stability and cooling speed. A comparable hybrid-metal mold system was investigated by [29], who demonstrated that integrating stainless steel inserts with aluminum substrates significantly enhanced thermal response without compromising the dimensional stability required for precision forming. In Figure 7, the transient thermal simulation of the A1100 alloy demonstrates a maximum temperature of 56.708°C concentrated on the upper surface of the frame, particularly at the central section where heat input occurs.

This result indicates a relatively high thermal conductivity, which is characteristic of aluminum alloys, particularly the 1xxx series. According to [29], aluminum A1100 possesses a thermal conductivity exceeding 220 W/m·K, which enables rapid heat diffusion across its structure, reducing localized temperature gradients. The uniform temperature distribution pattern observed in Figure 7 supports this finding, showing that the A1100 material efficiently disperses heat throughout the structural framework within a short transient period of 28 minutes. In contrast, Figure 8 illustrates the heat transfer behavior of stainless steel 304 (S304), which displays a maximum temperature of 56.002°C under the same boundary conditions. Although the difference in maximum temperature appears minimal, the thermal propagation pattern reveals a slower and more localized heat transfer process. The lower conductivity of S304, approximately 16 W/m·K as reported by Alsoufi and Bawazeer [32] , results in a steeper thermal gradient and delayed heat dissipation. This is evident from the more concentrated temperature zones near the heat source, suggesting that stainless steel retains heat for a longer duration compared to aluminum alloys.

When compared to prior works, the results are consistent with those of [30], who demonstrated that A1100 exhibits up to 85% faster transient heat transfer compared to austenitic stainless steels under equivalent convective conditions. Similarly, the study by [29] highlighted that aluminum's higher conductivity makes it more suitable for applications requiring rapid heat dissipation, such as food-grade processing equipment, where thermal stability and uniform heat distribution are essential to prevent localized overheating and maintain hygienic conditions. From the comparative analysis, it is evident that A1100 provides superior thermal performance in transient regimes, allowing faster equilibrium between hot and cold regions. Meanwhile, S304 offers better structural rigidity and corrosion resistance but at the expense of thermal responsiveness. The findings correspond well with the thermal performance index proposed by [38], which identified aluminum alloys as more energy-efficient for dynamic thermal environments, while stainless steels remain preferable for mechanical strength and chemical inertness in food-grade systems.

IV. CONCLUSIONS

The thermal simulation and statistical analysis conclusively demonstrate that material configuration significantly influences heat transfer characteristics in mold applications. The A1100-A1100 mold showed the fastest heat dissipation rate due to aluminum's high thermal conductivity; however, this also resulted in reduced heat retention and potential temperature non-uniformity at the molding interface. Conversely, the A1100-S304 hybrid mold displayed a slightly lower peak temperature but maintained a more uniform and controlled thermal gradient throughout the cycle. The ANOVA test results ($F = 12.83$, $p < 0.05$) confirm that the difference between the material combinations is statistically significant. The hybrid A1100–S304 configuration achieved superior overall thermal performance, balancing conductive efficiency and temperature stability. This outcome aligns with studies by [39],[40], which also emphasized the advantages of hybrid metal systems in achieving both energy efficiency and mechanical robustness. From an industrial perspective, the A1100–S304 mold combination offers the most practical solution for sustainable, food-safe thermal systems, reducing energy consumption while ensuring uniform product curing and minimizing deformation. The results highlight that hybrid metal engineering is a promising direction for next-generation biodegradable packaging molds, combining lightweight design, heat efficiency, and durability in a single system.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Agus Nuramal, Rama Dani Eka Putra, and Hendri Hestiawan.
Methodology: Agus Nuramal and Rama Dani Eka Putra.
Investigation: Agus Nuramal and Rama Dani Eka Putra.
Discussion of results: Agus Nuramal, Rama Dani Eka Putra and Hendri Hestiawan.
Writing – Original Draft: Rama Dani Eka Putra.
Writing – Review and Editing: Agus Nuramal and Rama Dani Eka Putra.
Resources: Rama Dani Eka Putra.
Supervision: Rama Dani Eka Putra and Hendri Hestiawan.
Approval of the final text: Agus Nuramal, Rama Dani Eka Putra and Hendri Hestiawan.

VI. REFERENCES

[1] C. Pothiraj, T. A. Gokul, K. R. Kumar, A. Ramasubramanian, A. Palanichamy, K. Venkatachalam, et al., "Vulnerability of microplastics on marine environment: A review," Ecol. Indic., vol. 155, no. July, 2023, doi: 10.1016/j.ecolind.2023.111058.

[2] M. Puttegowda, "Eco-friendly composites: exploring the potential of natural fiber reinforcement," Discov. Appl. Sci., vol. 7, no. 5, 2025, doi: 10.1007/s42452-025-06981-8.

[3] T. Eleutério, M. J. Trota, M. G. Meirelles, and H. C. Vasconcelos, "A Review of Natural Fibers: Classification, Composition, Extraction, Treatments, and Applications," Fibers, vol. 13, no. 9, pp. 1–39, 2025, doi: 10.3390/fib13090119.

[4] K. Z. M. A. Motaleb, A. Ahad, G. Laureckiene, and R. Milasius, "Innovative banana fiber nonwoven reinforced polymer composites: Pre-and post-treatment effects on physical and mechanical properties," Polymers (Basel)., vol. 13, no. 21, 2021, doi: 10.3390/polym13213744.

[5] T. A. Nguyen and T. H. Nguyen, "Study on Mechanical Properties of Banana Fiber-Reinforced Materials Poly (Lactic Acid) Composites," Int. J. Chem. Eng., vol. 2022, 2022, doi: 10.1155/2022/8485038.

[6] C. A. Ikedionu, F. Agbo Idoko, and O. J. Ojochogwu, "Design and Development of Biodegradable Polymers for Green Packaging Solutions," Int. J. Healthc. Sci., vol. 12, no. 2, pp. 84–110, 2024. https://doi.org/0.5281/zenodo.14793022

[7] A. I. Petaloti and D. S. Achilias, "The Development of Sustainable Biocomposite Materials Based on Poly(lactic acid) and Silverskin, a Coffee Industry By-Product, for Food Packaging Applications," Sustain., vol. 16, no. 12, 2024, doi: 10.3390/su16125075.

[8] A. Gaspar-Cunha, J. Melo, T. Marques, and A. Pontes, "A Review on Injection Molding: Conformal Cooling Channels, Modelling, Surrogate Models and Multi-Objective Optimization," Polymers (Basel)., vol. 17, no. 7, pp. 1–37, 2025, doi: 10.3390/polym17070919.

[9] E. J. Mahdi, H. F. Hussein, and F. R. Saeed, "Thermal performance comparison of aluminum and iron alloys in heat exchangers for solar water heating systems: Experimental study under Iraqi climate conditions," Next Mater., vol. 8, no. May, p. 100935, 2025, doi: 10.1016/j.nxmate.2025.100935.

[10] J. A. Sayyed, A. N. Nabeel, and A. A. Firdaus, "Exploring Thermal Conductivity: Experimental Insights Into Heat Transfer Properties Of Common Materials," Int. J. Environ. Sci., vol. 11, no. 11s, pp. 1142–1156, 2025, doi: 10.64252/5sxkbj63.

[11] M. Ilunga and A. Agarwal, "A Finite-Element-Analysis-Based Feasibility Study for," 2024.

[12] S. R. Shahriar, L. Jiang, J. Park, M. S. Islam, B. Perez, and X. Peng, "Experimental and FEA Simulations Using ANSYS on the Mechanical Properties of Laminated Object Manufacturing (LOM) 3D-Printed Woven Jute Fiber-Reinforced PLA Laminates," J. Manuf. Mater. Process., vol. 8, no. 4, 2024, doi: 10.3390/jmmp8040152.

[13] S. Bibi and M. Sajid, "Towards Cost Modelling for Rapid Prototyping and Tooling Technology-Based Investment Casting Process for Development of Low-Cost Dies," no. Cic, p. 6, 2025, doi: 10.3390/materproc2025023006.

[14] K. Sofias, Z. Kanetaki, C. Stergiou, A. Kantaros, S. Jacques, and T. Ganetsos, "Implementing CAD API Automated Processes in Engineering Design: A Case Study Approach," Appl. Sci., vol. 15, no. 14, pp. 1–26, 2025, doi: 10.3390/app15147692.

[15] D. Markus, A. Lorin, N. Thomas, and M. Sven, "Identifying an opportunistic method in design for manufacturing: An experimental study on successful a on the manufacturability and manufacturing effort of design concepts," Procedia CIRP, vol. 100, no. March, pp. 720–725, 2021, doi: 10.1016/j.procir.2021.05.090.

[16] P. Gurusamy, T. Satish, V. Mohanavel, A. Karthick, M. Ravichandran, O. Nasif, et al., "Finite Element Analysis of Temperature Distribution and Stress Behavior of Squeeze Pressure Composites," Adv. Mater. Sci. Eng., vol. 2021, 2021, doi: 10.1155/2021/8665674.

[17] S. Behseresht, Y. H. Park, A. Love, and O. A. Valdez Pastrana, "Application of Numerical Modeling and Finite Element Analysis in Fused Filament Fabrication: A Review," Materials (Basel)., vol. 17, no. 17, 2024, doi: 10.3390/ma17174185.
[18] J. Duan, P. Ma, G. Zhang, F. Gao, Y. Li, and Z. Wu, "Research on internal stress evolution mechanism in thermal processing of large precision machine tool components," Case Stud. Therm. Eng., vol. 75, no. June, p. 107038, 2025, doi: 10.1016/j.csite.2025.107038.

[19] A. Kuzmin, A. Ashori, P. Pantyukhov, Y. Zhou, L. Guan, and C. Hu, "Mechanical, thermal, and water absorption properties of HDPE/barley straw composites incorporating waste rubber," Sci. Rep., vol. 14, no. 1, pp. 1–11, 2024, doi: 10.1038/s41598-024-76337-6.

[20] Y. Kodali and Y. V. P. Kumar, "ANOVA-Based Variance Analysis in Smart Home Energy Consumption Data Using a Case Study of Darmstadt Smart City, Germany†," Eng. Proc., vol. 82, no. 1, pp. 1–8, 2024, doi: 10.3390/ecsa-11-20354.

[21] P. Luangpaiboon, W. Atthirawong, A. Hirunwat, and P. Aungkulanon, "Statistical learning-driven parameter tuning in injection molding using modified simplex method," Results Eng., vol. 27, no. May, p. 106690, 2025, doi: 10.1016/j.rineng.2025.106690.

[22] A. A. Firoozi, A. A. Firoozi, D. O. Oyejobi, S. Avudaiappan, and E. S. Flores, "Emerging trends in sustainable building materials: Technological innovations, enhanced performance, and future directions," Results Eng., vol. 24, no. November, p. 103521, 2024, doi: 10.1016/j.rineng.2024.103521.

[23] V. Oliver-Cuenca, V. Salaris, P. F. Muñoz-Gimena, A. Agüero, M. A. Peltzer, V. A. Montero, et al., "Bio-Based and Biodegradable Polymeric Materials for a Circular Economy," Polymers (Basel)., 2024.

[24] L. Quispe-Sanchez, R. Chuquilín-Goicochea, H. M. Figueroa-Avalos, S. G. Chavez, I. Yoplac, S. Mori, et al., "Biodegradable trays of cassava starch reinforced with Ceiba, coffee and cocoa fibers: a sustainable alternative to plastics," Appl. Food Res., vol. 5, no. 2, 2025, doi: 10.1016/j.afres.2025.101277.

[25] C. C. Kuo, J. Y. Xu, Y. J. Zhu, and C. H. Lee, "Effects of Different Mold Materials and Coolant Media on the Cooling Performance of Epoxy-Based Injection Molds," Polymers (Basel)., vol. 14, no. 2, 2022, doi: 10.3390/polym14020280.

[26] J. L. Mullo, I. L. Fé-Perdomo, J. Ramos-Grez, Á. F. M. Romero, A. Ramírez-Albán, M. Yarad-Jácome, et al., "Predicting the Relative Density of Stainless Steel and Aluminum Alloys Manufactured by L-PBF Using Machine Learning," J. Manuf. Mater. Process., vol. 9, no. 6, pp. 1–16, 2025, doi: 10.3390/jmmp9060185.

[27] J. Garavito, C. P. Peña-Venegas, and D. A. Castellanos, "Production of Starch-Based Flexible Food Packaging in Developing Countries: Analysis of the Processes, Challenges, and Requirements," Foods, vol. 13, no. 24, 2024, doi: 10.3390/foods13244096.

[28] R. M. Badr, E. Amine, B. Fatima, R. Mohammed, D. Meryiem, M. Hicham, et al., "Sustainable Building Materials: Optimization and Performance Analysis of Plaster/Wood Shavings Composites for Thermal Insulation," J. Compos. Sci., vol. 9, no. 6, pp. 1–21, 2025, doi: 10.3390/jcs9060289.

[29] A. Zhang and Y. Li, "Thermal Conductivity of Aluminum Alloys—A Review," Materials (Basel)., vol. 16, no. 8, 2023, doi: 10.3390/ma16082972.

[30] W. Ye, D. Jamshideasli, and J. M. Khodadadi, "Improved Performance of Latent Heat Energy Storage Systems in Response to Utilization of High Thermal Conductivity Fins," Energies, vol. 16, no. 3, 2023, doi: 10.3390/en16031277.

[31] Z. T. Al-Sharify, S. Z. Al-Najjar, Z. A. Naser, Z. A. I. Alsherfy, and H. Onyeaka, "The Impact of Fluid Flow on Microbial Growth and Distribution in Food Processing Systems," Foods, vol. 14, no. 3, pp. 1–24, 2025, doi: 10.3390/foods14030401.

[32] M. S. Alsoufi and S. A. Bawazeer, "Predictive Modeling of Surface Integrity and Material Removal Rate in Computer Numerical Control Machining: Effects of Thermal Conductivity and Hardness," Materials (Basel)., vol. 18, no. 7, 2025, doi: 10.3390/ma18071557.

[33] W. Tang, K. Zhou, Z. Li, L. Xin, D. Huang, F. Zhan, et al., "Differential Corrosion Behavior of High-Aluminum 304 Stainless Steel in Molten Nitrate Salts: The Roles of Rolling and Heat Treatment," Materials (Basel)., vol. 18, no. 19, p. 4513, 2025, doi: 10.3390/ma18194513.

[34] R. Lv, C. Yin, B. Bai, W. Yang, and Z. Zhou, "The Microstructure and Mechanical Properties of a 15-6 PH Stainless Steel with Improved Thermal Aging Embrittlement Resistance," Materials (Basel)., vol. 17, no. 5, 2024, doi: 10.3390/ma17051179.

[35] E. Fernandez, M. Edeleva, R. Fiorio, L. Cardon, and D. R. D'Hooge, "Increasing the Sustainability of the Hybrid Mold Technique through Combined Insert Polymeric Material and Additive Manufacturing Method Design," Sustain., vol. 14, no. 2, 2022, doi: 10.3390/su14020877.

[36] R. Hussin, S. Sharif, M. Nabiałek, S. Z. A. Rahim, M. T. M. Khushairi, M. A. Shuaimi, et al., "Hybrid mold: Comparative study of rapid and hard tooling for injection molding application using metal epoxy composite (MEC)," Materials (Basel)., vol. 14, no. 3, 2021, doi: 10.3390/ma14030665.

[37] S. Cho, J. Lee, S. Shin, D. Lee, M. Kim, H. Kwon, et al., "Enhancing high-temperature properties of stainless steel composite with titanium carbide reinforcement: a study on coefficient of thermal expansion, thermal conductivity, and strength," J. Mater. Res. Technol., vol. 25, pp. 7241–7253, 2023, doi: 10.1016/j.jmrt.2023.07.162.

[38] K. F. Lunn and D. Apelian, "Thermal and Electrical Conductivity of Aluminum Alloys: Fundamentals, structure-property relationships, and pathways to enhance conductivity," Mater. Sci. Eng. A, vol. 924, no. December 2024, p. 147766, 2025, doi: 10.1016/j.msea.2024.147766.

[39] D. Nayak, S. A. Rao, S. L. Gaonkar, and P. Preethi Kumari, Hybrid metal matrix composite: a comprehensive review on its fabrication, corrosion behaviour and inhibition, vol. 7, no. 8. Springer International Publishing, 2025. doi: 10.1007/s42452-025-07288-4.

[40] J. P. M. Pragana, R. F. V. Sampaio, I. M. F. Bragança, C. M. A. Silva, and P. A. F. Martins, "Hybrid metal additive manufacturing: A state–of–the-art review," Adv. Ind. Manuf. Eng., vol. 2, no. January, p. 100032, 2021, doi: 10.1016/j.aime.2021.100032.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 296-304. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2899

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

DESIGN, FABRICATION, AND EVALUATION OF A LOW-COST AND MOBILE FISH SMOKEHOUSE FOR SMOKED SOFT-BONED MILKFISH

Ervin F. Papa*[1], Jesus Paolo B. Cullado[2], Quennie M. Bilugan[3] and Ma. Elizabeth E. Bravo[4]

[1, 2, 3, 4]Cavite State University Naic, Bucana, Naic, Cavite, Philippines.

[1]http://orcid.org/0000-0003-0467-0610, [2]http://orcid.org/0009-0005-8762-9307, [3]https://orcid.org/0000-0003-2613-2426, [4]https://orcid.org/0009-0008-0003-2439

Email: *ervin.papa@cvsu.edu.ph, jesuspaolo.cullado@cvsu.edu.ph, quennie.bilugan@cvsu.edu.ph, ma.elizabeth.esmeralda@cvsu.edu.ph

ARTICLE INFO

Article History
Received: November 7, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Mobile smokehouse,
Fish processing,
Smoked fish,
Sustainable technology,
Economic viability.

ABSTRACT

This study details the design, fabrication, and evaluation of a low-cost (PhP 20,867.63), mobile fish smokehouse to support livelihood programs in Naic, Cavite, Philippines. The unit, built from a recycled 200-liter drum, was tested using a 2 x 2 factorial experiment. Results showed smoking time significantly impacted moisture removal ($p < 0.01$), achieving an average rate of 52.63 ± 2.86 %/hr. A cost-benefit analysis demonstrated high financial viability, with a 167% ROI and a 7.19-month payback period. A key limitation was identified: comparative testing showed the unit had a significantly lower moisture removal rate ($p=0.004$) and received lower sensory acceptability scores ($p=0.045$) than a conventional smokehouse, likely due to suboptimal thermal efficiency and smoke distribution. Despite these performance trade-offs, the smokehouse proves to be a financially viable and portable solution for community-level extension and supplementary livelihood.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Fish is one of the most popular and widely consumed protein foods in the world. This aquatic resource includes several species that people commonly buy for consumption. Fresh fish are in demand among people of all social groups. High production that causes surplus supplies and spoilage is one of the key reasons for developing post-harvest products or practicing fish preservation. Fish smoking is one of the most popular preservation techniques in developing countries due to its simplicity and low operational inputs, holding significant economic importance for the seafood sector [1]. This process not only mitigates high production leading to surplus supplies and spoilage but also extends the market reach of fish products. Smoking is highly adaptable to small-scale and community operations, even in coastal regions and inland communities close to water bodies. This method, a cornerstone of coastal livelihoods in the Philippines, often faces constraints due to available technology [2].

Fish processing is a vital source of supplementary income in the province of Cavite, particularly in coastal municipalities like Naic. The Cavite State University Naic (CvSU Naic) research and extension unit is mandated to convert research into sustainable extension projects that will help improve the lives of the people in its service areas, aligning with national fishery objectives to upgrade post-harvest technologies. A continuous extension project that promotes smoked soft-boned milkfish technology has identified a pressing need for an innovative processing solution. This project specifically targets groups like the Naic Solo Parents Association to provide livelihood and increase income. However, significant socioeconomic barriers hinder the adoption and scaling of this technology. High initial capital—conventional smokehouses, even simplified local models, require a substantial initial investment [3], [4]. Logistical constraints & lack of portability—existing smokehouse designs are typically bulky, stationary, and non-modular [2]. This lack of portability makes it impractical for extension workers to conduct hands-on training across different barangays (villages) in Naic, and it also prevents beneficiaries from relocating their production to match seasonal fish landings or access different local markets.

Livelihood barriers—the combination of high cost and immobility means that traditional smoking technology is not accessible as a supplementary livelihood, which requires low-risk, low-cost, and flexible solutions. This technology gap directly impedes sustainable livelihood improvement and economic empowerment at the community level. Addressing this gap aligns with broader development goals, including Sustainable Development Goal 8 (Decent Work and Economic Growth) by fostering local entrepreneurship and SDG 2 (Zero Hunger) by improving food preservation, adding value to local products, and reducing post-harvest losses [5]. Therefore, this research directly addresses this identified gap by focusing on the innovative design, efficient fabrication, and rigorous evaluation of a novel low-cost and mobile fish smokehouse. This unit was made to help small-scale fish processors to overcome the financial and logistical problems they face. It provides a practical, portable, and affordable way to make high-quality smoked soft-boned milkfish at the community level.While various "improved" smoking kilns, such as the Chorkor oven, FTT (FAO-Thiaroye Technique), and other locally modified cabinet-type smokers, have been developed, many still face challenges of high fabrication cost, lack of portability, or complex operation [2], [6-8].

Conventional models are often stationary, limiting their use to a central processing facility [3], [9], [10]. The primary novelty of the smokehouse proposed in this study lies not in a new smoking chemistry but in its synthesis of critical design features tailored for community extension and small-scale entrepreneurship: (1) affordability—the design prioritizes the use of locally sourced, low-cost, and recycled materials, primarily a 200-liter food-grade steel drum, achieving a final fabrication cost of PhP 20,867.63—well below many standardized 'improved' models; (2) mobility—unlike stationary kilns, this unit is integrated onto a wheeled frame, transforming it from a fixed piece of equipment into a portable technology-demonstration tool, this directly addresses the logistical challenges of extension work, allowing the smokehouse to be brought to different barangays for on-site training; (3) specific applicability—the combination of low cost and mobility makes it a highly applicable solution for supplementary livelihood programs, it lowers the barrier to entry for beneficiaries (e.g., Naic Solo Parents Association) and provides a flexible asset that can be shared or moved as needed. This design explicitly prioritizes financial accessibility and logistical flexibility over achieving maximum thermal efficiency or throughput, differentiating its application from larger, stationary kilns designed for established commercial operations [11], [10].

II. MATERIALS AND METHODS

II.1 DESIGN AND FABRICATION OF THE MOBILE SMOKEHOUSE

The low-cost, mobile fish smokehouse was designed as a vertical, indirect smoking system to prevent direct contact between the fish and fire/ash. The design prioritized two key factors: extreme affordability and portability, differentiating it from high-throughput industrial kilns [4], [12]. The main smoking chamber was fabricated from a readily available, recycled 200-liter food-grade steel drum. The total fabrication cost was documented as PHP 20,867.63. Figure 1 shows the major components, and Figure 2 displays the finalized fabricated unit.

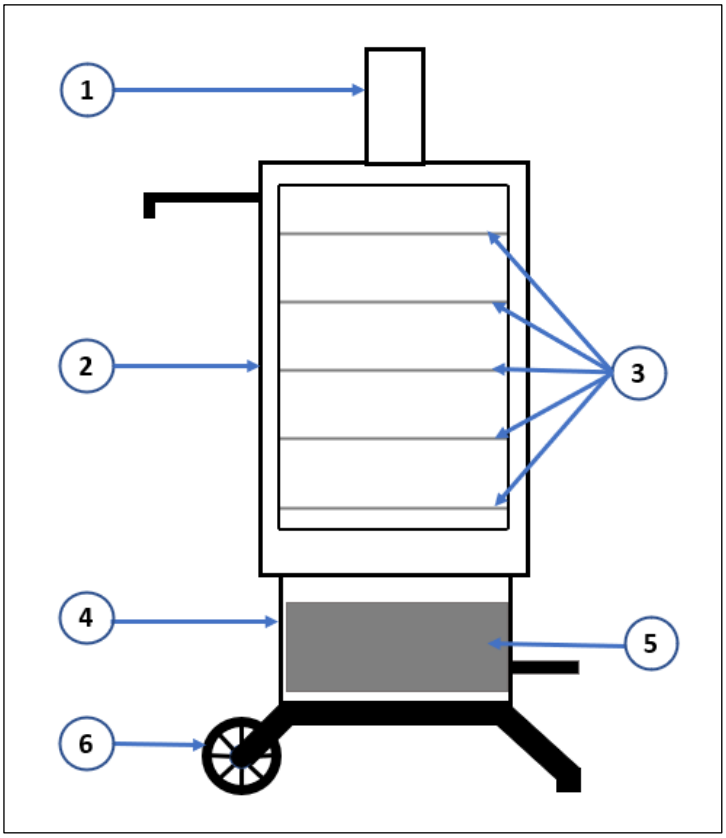


Figure 1: The major components of the low-cost fish smokehouse include: (1) chimney; (2) smoking chamber; (3) smoking trays; (4) heating chamber; (5) charcoal pot; and (6) mobility frame.
Source: Authors (2026).

Figure 3 demonstrates the functionality of the mobility features incorporated into the design and fabrication of the fish smokehouse. It demonstrates that the wheels and push handle allow for the relatively easy movement of the assembled unit by a single person, highlighting the successful implementation of the "mobile" design consideration. This portability would be advantageous for transporting the smokehouse to different locations for use, cleaning, or storage, potentially increasing its practicality for small-scale fish processors.



Figure 2: The low-cost and mobile fish smokehouse.
Source: Authors (2026).



Figure 3: Mobility testing of the fabricated fish smokehouse.
Source: Authors (2026).

Table 1 presented the established specifications of the major components of the low-cost and mobile fish smokehouse. The material selections for each component of the low-cost and mobile fish smokehouse were considered in its design considerations of low cost, local availability, food-grade suitability, ease of operation, and portability. The simple pipe design of the chimney facilitated unobstructed airflow, and the adjustable damper provided basic control over smoke and heat regulation. The 200-liter food-grade steel drum indicates an emphasis on low cost and the potential for using readily available recycled materials. The notable features included a hinged door with a latch. The five levels of tray support allowed for smoking multiple layers of fish simultaneously, increasing capacity. The hinged door with a latch ensured easy access for loading and unloading while maintaining a sealed environment during smoking. The fabrication of the heating chamber utilizing steel plate with heat-resistance paint reflected the ability to withstand the direct heat source. The key features of the heating chamber included its separate location below the smoking chamber and smoke vents positioned between the chambers. The charcoal pot was tailored to fit within the heating chamber, ensuring efficient heat transfer and heat resistance from charcoal.

Table 1: Specifications of the major components of the low-cost and mobile fish smokehouse.

MAJOR COMPONENTS	SPECIFICATION		
	Material	**Dimension**	**Features**
1. Chimney	● rounded steel plate ● heat resistance paint	● 125 mm diameter ● 250 mm height ● 2 mm thick	● adjustable damper at the top
2. Smoking chamber	● 200-liter food-grade steel drum ● heat resistance paint	● 584 mm diameter ● 901.7 mm height ● 2 mm thick ● 150 mm vertical distance of tray support	● hinged door with latch ● five levels of tray support (stainless)
3. Smoking trays	● 304-stainless steel rod	● 500 mm diameter	● with two handles
4. Heating chamber	● steel plate ● heat resistance paint	● 500 mm width ● 300 mm length ● 300 mm height ● 2 mm thick	● separate chamber below the smoking chamber ● smoke vents in between chambers
5. Charcoal pot	● steel plate ● heat resistance paint ● steel pipes	● 450 mm width ● 250 mm length ● 250 mm height ● 2 mm thick	● fabricated to fit inside the heating chamber ● with two handles
6. Mobility frame	● angular steel bars ● steel rod ● rubber rim and bearings (wheel)	● various lengths on the parts for the mobility smokehouse	● with two wheels ● base frame ● stand ● push handle

Source: Authors (2026).

II.2 EXPERIMENTAL DESIGN AND PERFORMANCE EVALUATION

The performance evaluation of the mobile smokehouse focused on its efficiency in removing moisture, a critical parameter for quality and shelf life [13]. Preliminary testing was conducted at the CvSU Naic Mabolo Annex to evaluate the operative conditions for processing smoked soft-boned milkfish in the fabricated fish smokehouse. The preliminary testing of the low-cost and mobile fish smokehouse involved the observation and recording of the duration required to achieve the moisture removal rate (%/hr) and the intended golden-brown color of the smoked fish. The study employed a 2x2 factorial experiment in a Completely Randomized Design (CRD) to investigate the effects of smoking time and loading capacity on the moisture removal rate of soft-boned milkfish processed in the fabricated low-cost fish smokehouse. Two independent variables were examined: smoking time at two levels (30 minutes and 60 minutes) and loading capacity at two levels (10 kg and 15 kg). The experimental unit consisted of a batch of five (5) individual soft-boned milkfish samples.

These samples were grouped according to the four combined levels of the independent variables, creating four distinct experimental treatments. The assignment of experimental treatments to different batches of fish was randomized. Furthermore, each of the four treatment combinations was replicated three (3) times. This resulted in a total of twelve (12) experimental units (batches of fish) being processed and analyzed. The dependent variable measured in this experiment was the average moisture removal rate (%/hr) for each batch of smoked soft-boned milkfish under the specified treatment conditions. A protocol was established to facilitate the production of smoked soft-boned milkfish using a fabricated fish smokehouse. The main phases of the protocol to facilitate the production of smoked soft-boned milkfish using a mobile and low-cost fish smokehouse are illustrated in Figure 4. The subsequent processes are as follows: (1) freezing, (2) thawing, (3) cleansing, (4) soaking in brine solution, (5) pressure cooking, (6) drying, (7) smoking, (8) packaging, and (9) storing.

II.3 SENSORY EVALUATION

The finished smoked milkfish products from both the mobile unit and the conventional smokehouse were subjected to a consumer acceptability test. A preference or acceptance test, utilizing a 5-point hedonic scale, was adopted as the sensory testing method [14], [15] Thirty (30) participants were recruited for the taste tests, comprising ten (10) experienced smoked soft-boned processors, ten (10) instructors/professors (five from the Fisheries and Aquatic Sciences Department (FASD) and five from the Management Department (MD), and ten (10) Bachelor of Science in Fisheries and Aquatic Sciences students. The organoleptic parameters assessed during the taste test included color, taste, aroma, texture, and general acceptability. For each parameter, participants indicated their preference using the 5-point hedonic scale. The taste tests were carried out in the Conference Room, CvSU Naic. Prior to the evaluation, a brief orientation was conducted by the study staff to explain the procedure, duration, and materials involved. The evaluation form was also thoroughly explained to each participant. Each participant was assigned a designated taste test booth equipped with the test samples, rating forms, a pencil with an eraser, clean drinking water, a comfortable adjustable chair, a spoon, a fork, and a small knife.

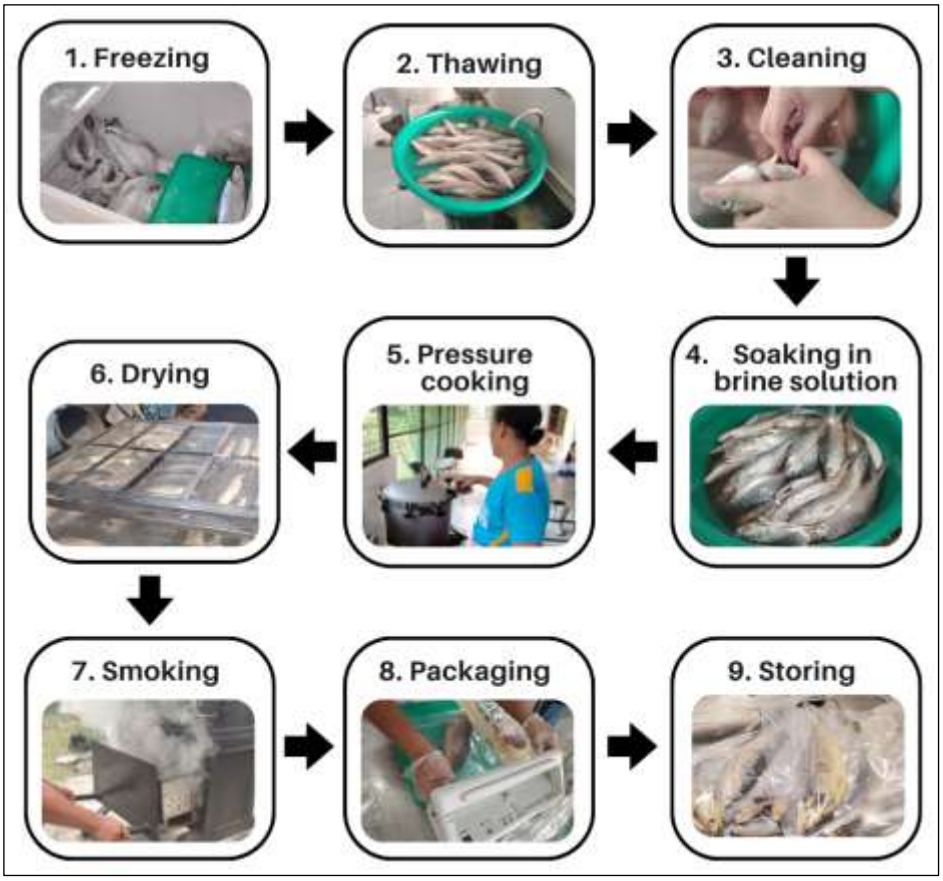


Figure 4: Main phases of the protocol for smoked soft-boned milkfish production.
Source: Authors (2026).

II.4 STATISTICAL ANALYSIS AND COST-BENEFIT ANALYSIS

The data collected from the various stages of the study were subjected to appropriate statistical analyses to address the research objectives. For objectives that focused on the design, fabrication, and evaluation of the newly developed low-cost fish smokehouse, descriptive statistics (means and standard deviations) were calculated to summarize the moisture removal rate under different loading capacities (10 kg and 15 kg) and smoking times (30 minutes and 60 minutes). A two-way analysis of variance (ANOVA) was employed to determine the significant effects of loading capacity and smoking time, as well as their interaction, on the moisture removal rate. The significance level was set at $p < 0.05$. For objectives that involved a comparative performance evaluation of the existing and newly developed fish smokehouses, the average moisture removal rate of the existing smokehouse was calculated using descriptive statistics. An independent samples t-test was used to see how the two smokehouses' moisture removal rates compared. The p-value for significance was set at 0.05. The sensory acceptability data, obtained from the preference tests, were analyzed using descriptive statistics for each sensory attribute for the smoked soft-boned milkfish produced by both the newly developed and the existing smokehouses. To determine if there were significant differences in the overall acceptability between the products of the two smokehouses, an independent samples t-test was performed, with the significance level set at $p < 0.05$. A simple cost analysis was performed for the newly constructed smokehouse. This analysis involved the calculation of the payback period (PBP), breakeven point (BEP), benefit-cost ratio (BCR), and return on investment (ROI) based on the estimated costs of materials, fabrication, and potential income from the smoked fish production.

III. RESULTS AND DISCUSSIONS

III.1 OPERATING CONDITIONS TO PROCESS SMOKED SOFT-BONED MILKFISH

Figure 5 illustrates the moisture removal rate and change in the color of the smoked soft-boned milkfish over time. The findings indicated that the newly developed low-cost smokehouse could achieve the desired "golden brown" color for smoked soft-boned milkfish within a relatively short smoking time of approximately 30 minutes. This suggested efficient heat and smoke transfer within the smoking chamber. Table 2 shows how the loading capacity (10 kg and 15 kg) and smoking time (30 minutes and 60 minutes) affect the moisture removal rate (%/hr) of the smoked soft-boned milkfish. The results showed that, regardless of the loading capacity, the average moisture removal rate was notably higher for the 30-minute smoking time (52.63 ± 2.86 %/hr) compared to the 60-minute smoking time (27.42 ± 2.21 %/hr). The total average moisture removal rates for the 10 kg loading capacity (40.81 ± 14.49 %/hr) and the 15 kg loading capacity (39.24 ± 13.53 %/hr) were relatively similar. The significantly higher moisture removal rate observed at the 30-minute smoking time compared to the 60-minute smoking time, irrespective of the loading capacity, suggested that the initial stage of smoking was more effective in removing moisture.

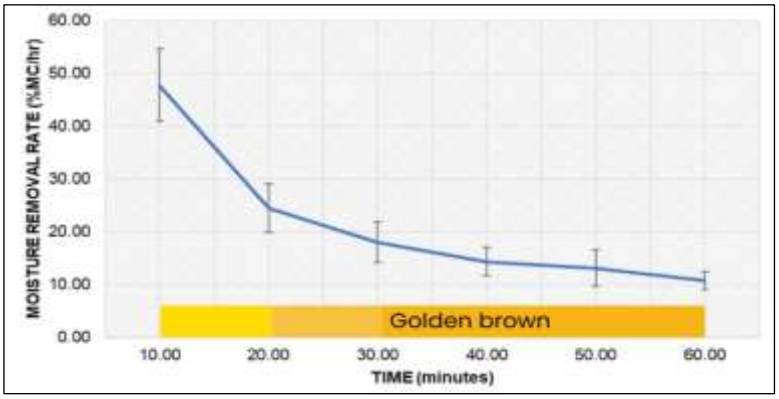


Figure 5: The change in the color and moisture removal rate of the smoked soft-boned milkfish over time.
Source: Authors (2026).

Table 2: Effects of loading capacity and smoking time on the moisture removal rate of smoked soft-boned Milkfish.

LOADING CAPACITY (kg)	SMOKING TIME (min)	MOISTURE REMOVAL RATE (%/hr) Mean	± S.D.	BATCH (Samples)
10.00	30.00	53.76	± 3.86	3 (15)
	60.00	27.86	± 2.64	3 (15)
	Total	40.81	± 14.49	6 (30)
15.00	30.00	51.50	± 1.30	3 (15)
	60.00	26.97	± 2.15	3 (15)
	Total	39.24	± 13.53	6 (30)
Total	30.00	52.63	± 2.86	6 (30)
	60.00	27.42	± 2.21	6 (30)
	Total	40.02	± 13.39	12 (60)

Source: Authors (2026).

III.2 ACCEPTABILITY OF SMOKED SOFT-BONED MILKFISH PRODUCED USING THE EXISTING AND NEWLY DEVELOPED FISH SMOKEHOUSE

Table 3 presents the sensory acceptability scores for smoked soft-boned milkfish produced using both the newly developed and the existing smokehouses, evaluated across five attributes. The scores were based on a 5-point hedonic scale, where 5 represented "Like Extremely" or the most desirable attribute, and 1 represented "Dislike Extremely" or the least desirable.

Table 3: Sensory acceptability scores for smoked soft-boned milkfish produced using both the newly developed and the existing smokehouses.

Attributes	Acceptability	
	Newly developed smokehouse	Existing smokehouse
Color	3.33 ± 0.96 (Slightly uneven)	4.58 ± 0.61 (Very appealing)
Taste	3.94 ± 0.70 (Balanced)	4.42 ± 0.71 (Delicious)
Aroma	4.09 ± 0.63 (Pleasant smoky aroma)	4.42 ± 0.56 (Very Appealing)
Texture	4.06 ± 0.66 (Firm, moist and flaky texture)	4.03 ± 0.77 (Firm, moist and flaky texture)
General acceptability	4.03 ± 0.82 (Like Moderately)	4.42 ± 0.75 (Like Extremely)

Source: Authors (2026).

For color, the smoked milkfish from the existing smokehouse received a significantly higher mean score of 4.58 ± 0.61, corresponding to "Very Appealing," compared to the newly developed smokehouse, which had a mean score of 3.33 ± 0.96, categorized as "Slightly uneven." In terms of taste, the existing smokehouse also yielded a higher mean score of 4.42 ± 0.71 ("Delicious") compared to the newly developed smokehouse, which received a mean score of 3.94 ± 0.70 ("Balanced"). Similarly, the aroma of the smoked milkfish from the existing smokehouse was rated higher with a mean of 4.42 ± 0.56 ("Very Appealing") compared to the newly developed smokehouse, which had a mean of 4.09 ± 0.63 ("Pleasant smoky aroma"). Regarding texture/mouthfeel, the mean scores were comparable between the two smokehouses. The existing smokehouse received a score of 4.03 ± 0.77 ("Firm, moist, and flaky texture"), while the newly developed smokehouse was rated at 4.06 ± 0.66 ("Firm, moist, and flaky texture"). Finally, for overall acceptability, the smoked soft-boned milkfish produced by the existing smokehouse received a significantly higher mean score of 4.42 ± 0.75 ("Like Extremely") compared to the newly developed smokehouse, which had a mean score of 4.03 ± 0.82 ("Like Moderately").

III.3 DIFFERENCE IN THE MOISTURE REMOVAL RATE (%/HR) BETWEEN THE EXISTING AND NEWLY DEVELOPED FISH SMOKEHOUSE

The newly developed smokehouse exhibited a mean moisture removal rate of 53.76 ± 3.86 %/hr, while the existing smokehouse had a mean rate of 59.13 ± 3.53 %/hr. There was a statistically significant difference in the moisture removal rate between the two smokehouses. Specifically, the average rate of moisture removal from the existing smokehouse (59.13 ± 3.53 %/hr) was significantly higher than that of the newly developed smokehouse (53.76 ± 3.86 %/hr). The statistical analysis revealed a significant difference in the efficiency of moisture removal between the newly developed and the existing fish smokehouses. Results suggested the existing smokehouse was more effective at reducing the moisture content of the fish under a loading capacity of 10 kg and a smoking time of 30 minutes (table 4).

Table 4: Independent samples t-test results comparing the average moisture removal rate.

Smokehouses	Average Moisture Removal Rate (%/hr)	t-statistics	df	p-value
	Mean ± S.D.			
Newly developed	53.76 ± 3.86	-3.210	25	0.004
Existing	59.13 ± 3.53			

Source: Authors (2026).

III.4 DIFFERENCE IN THE OVERALL ACCEPTABILITY OF THE EXISTING AND NEWLY DEVELOPED FISH SMOKEHOUSE'S SMOKED SOFT-BONED MILKFISH PRODUCTS

The mean overall acceptability score for the milkfish from the newly developed smokehouse was 4.03 ± 0.82 ("Like Moderately"), while the milkfish from the existing smokehouse had a mean score of 4.42 ± 0.75 ("Like Extremely"). The statistical analysis revealed a significant difference in the overall sensory acceptability of the smoked soft-boned milkfish produced by the two types of smokehouses. The panelists, comprising experienced smoked fish processors, fisheries experts, and students, exhibited a significantly higher preference for the product originating from the existing smokehouse compared to the product from the newly developed one.

Table 5: Independent samples t-test comparing the overall acceptability.

Smokehouses	Overall acceptability Mean ± S.D.	t-statistics	df	p-value
Newly developed	4.03 ± 0.82 (Like Moderately)	-2.049	64	0.045
Existing	4.42 ± 0.75 (Like Extremely)			

Source: Authors (2026).

III.5 SIMPLE COST ANALYSIS

A simple cost analysis was done to present the economic viability of fabricating the low-cost mobile smokehouse. The total cost of constructing one unit of the developed smokehouse is PHP 20,867.63. Table 6 outlines the key financial assumptions used in the cost analysis. These assumptions provide the foundation for understanding the economic viability of the low-cost fish smokehouse. Table 7 presents the results of the cost analysis, derived from the assumptions in Table 6. These results provide key metrics for evaluating the financial feasibility of the smokehouse.

Table 6: Basic assumptions for the simple cost analysis of the smokehouse.

Basic Assumptions	Value
Initial capital investment (IC) (PHP)	20,867.63
Useful life (years)	5
Estimated salvage value, 10% IC (PHP)	2,086.76
Repair and maintenance, 5% per year (PHP)	1,043.38
Cost of goods sold per batch (PHP)	4,350
Smokehouse capacity per batch (kg)	15
Number of batches per year	24
Selling price of smoked soft-boned bangus per kg (PHP)	400/kg
Revenue per batch (PHP)	6,000
Annual depreciation (PHP)	3,756.17
Annual gross profit (PHP)	39,600
Annual net profit (PHP)	34,800.45

Source: Authors (2026).

The simple cost analysis reveals promising economic viability for the low-cost fish smokehouse. The payback period of approximately 7.19 months indicates a rapid recovery of the initial investment of PHP 20,867.63. The substantial annual net profit of PHP 34,800.45 primarily contributes to this swift return. The breakeven point analysis demonstrates that the smokehouse operation will start generating profit after producing and selling around 13 batches of smoked fish, with a corresponding revenue of PHP 75,900. Based on the assumptions, the business is expected to make more money than it costs to run in the first year. The benefit-cost ratio (BCR) of 1.32 signifies a highly favorable return. For every peso invested in the fish smokehouse, approximately PHP 1.32 in benefits (revenue) is generated. Since the BCR exceeds 1, the project is considered economically viable and generates more benefits than costs.

Furthermore, the return on investment (ROI) of 167% indicates that the smokehouse is expected to yield a very high return relative to the initial capital invested. This suggests that the low-cost fish smokehouse has the potential to be a financially attractive venture.

Table 7: Simple cost analysis results.

Parameters	Value
Payback period (PBP)	
PBP, months	7.19
PBP, years	0.599
Breakeven Point (BEP)	
BEP, batches	12.65 ≈ 13
BEP, PHP	75,900.00
Benefit-Cost Ratio	1.32
Return on Investment, % (ROI)	167

Source: Authors (2026).

When compared to other low-cost smoking devices, the findings of this study are highly consistent with the trade-offs observed in technology scaling. For instance, the PHP 20,867.63 fabrication cost (approx. $350 USD) is extremely competitive. Bartolome et al. [4] reported an automated village-level machine with a higher capacity but a significantly higher cost of PHP 170,000.00. Other locally modified drum-type kilns have shown similar challenges in balancing initial investment, portability, and performance [9]. This comparison highlights that while our mobile unit trades some thermal performance (lower MRR and sensory scores) for a drastic reduction in cost and an increase in mobility, its financial viability (167% ROI) presents a unique and necessary advantage for the target socioeconomic context. The novelty lies not in technological perfection, but in optimizing the design for accessibility and extension work, thereby lowering the barrier to entry for marginalized groups in coastal communities.

III.6 DESIGN LIMITATIONS

A significant finding of this study was the trade-off between the primary goals of low cost and mobility and the resulting product quality. The sensory evaluation revealed that while the smoked milkfish from the mobile unit was acceptable (rated "Like Moderately"), it scored significantly lower in overall acceptability, color, taste, and aroma compared to the product from the existing, conventional smokehouse. This outcome is likely attributable to several key technical factors inherent in the prototype's design. Airflow Dynamics and Smoke Distribution—the mobile unit employs an indirect smoking method, where the smoke generator is separate from the main 200-liter drum chamber. While this design (shown in Figure 1) successfully separates the heat source from the food, it relies on natural convection to carry smoke to the fish. Without optimized baffles or a fan, the smoke and heat likely flowed unevenly through the cylindrical chamber. This hypothesis is strongly supported by the "slightly uneven" color noted in the sensory results (Table 7). Conventional, larger kilns often have a more optimized geometry or baffling systems to ensure uniform smoke contact [2], [11]. Smoke Density and Thermal Efficiency—in an indirect system, the smoke must travel from the heating chamber to the smoking drum.

This travel path can allow the smoke to cool, which may cause some of the "heavy" phenolic compounds responsible for desirable, rich smoke flavor and aroma to condense before reaching the fish [15]. Furthermore, the design's focus on low-cost materials (a single-walled steel drum) results in poor heat retention (low thermal efficiency) compared to insulated or thicker-walled kilns [16], [8]. The consequent lower, and potentially less stable, temperature profile within the smoking chamber could inhibit the Maillard reactions responsible for developing a deep, uniform golden-brown color, leading to the "slightly pale" appearance. Implications for Future Design—These results are crucial for guiding future improvements, as suggested by the reviewer. The lower sensory scores are not a failure of the concept but a clear indication of a performance trade-off. Future iterations should focus on optimizing the smoke-heat delivery system. As noted in the recommendations, integrating a simple baffling system, improving the insulation of the drum (e.g., a "drum-in-drum" design or external clay insulation), or even incorporating a small, low-cost solar-powered fan to regulate airflow could significantly improve smoke density and thermal distribution. These enhancements would directly address the uneven color and milder flavor profile, bridge the quality gap while maintaining the core benefits of low cost and mobility.

V. CONCLUSIONS

The results indicate that the low-cost and mobile fish smokehouse successfully met the basic design objectives of portability, affordability, and food safety. Its construction using locally sourced materials and its integration of indirect smoking principles provide a practical and replicable model for small-scale applications. Nevertheless, the evaluation revealed that the moisture removal efficiency and sensory quality of the smoked products were significantly lower than those produced by the existing conventional smokehouse. These findings highlight the need for further design refinements, particularly in improving thermal efficiency and smoke distribution within the chamber. Integrate a simple, low-cost baffling system within the smoking chamber to ensure a more uniform distribution of heat and smoke, which would directly address the "slightly uneven" color noted by panelists. Enhance the insulation of the smoking chamber. A low-cost "drum-in-drum" design or an outside layer of insulating material could be used to keep the heat in and keep the temperature inside stable. Explore the integration of a small, low-cost, solar-powered fan to create a forced-draft system. This would provide better control over smoke density and airflow, potentially improving both the moisture removal rate and the sensory profile. Future design iterations should employ advanced analysis tools to improve rigor. Integrating thermal imaging can visualize heat distribution in real-time, while Computational Fluid Dynamics (CFD) simulation can be used to model and optimize internal airflow and smoke distribution before fabrication.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Ervin F. Papa and Jesus Paolo B. Cullado
Methodology: Ervin F. Papa, Jesus Paolo B. Cullado, Quennie M. Bilugan and Ma. Elizabeth E. Bravo
Investigation: Ervin F. Papa, Jesus Paolo B. Cullado, Quennie M. Bilugan and Ma. Elizabeth E. Bravo
Discussion of results: Ervin F. Papa, Jesus Paolo B. Cullado, Quennie M. Bilugan and Ma. Elizabeth E. Bravo
Writing – Original Draft: Ervin F. Papa, Jesus Paolo B. Cullado, Quennie M. Bilugan and Ma. Elizabeth E. Bravo
Writing – Review and Editing: Ervin F. Papa
Resources: Jesus Paolo B. Cullado and Quennie M. Bilugan
Supervision: Ervin F. Papa
Approval of the final text: Ervin F. Papa

VII. ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of several individuals and institutions whose contributions were essential to the completion of this study. Ms. Linda Gumahis and Mr. Joel C. Credito are thanked for their invaluable assistance during the experiment. The Research Development and Extension Services (RDES) and the Knowledge Management Center (KMC) are recognized for their guidance and support, including assistance with intellectual property applications. The authors also extend their gratitude to Cavite State University for providing the necessary funding for this research, and to Prof. Armina B. Pugay, the Campus Administrator of CvSU Naic, for her continuous support throughout the project.

VIII. REFERENCES

[1] F. F. Olayemi, R. A. Olayinka, O. A. Oyelese, S. N. Oyewole, and M. A. Omodara, "Effective fish smoking kiln for developing country," Int. J. Sci. Eng. Res., vol. 4, no. 1, pp. 1–7, Jan. 2013.

[2] E. K. Asamoah et al., "Comparison of the efficiency of improved and traditional fish smoking kilns and their effects on smoked fish quality in Ghana," Journal of the Science of Food and Agriculture, vol. 105, no. 5, pp. 2923–2930, 2025. doi: 10.1002/jsfa.14060.

[3] K. Belichovska, D. Belichovska, and Z. Pejkovski, "Smoke and smoked fish production," Meat Technology, vol. 60, no. 1, pp. 37–43, Jan. 2019. doi: 10.18485/meattech.2019.60.1.6.

[4] G. J. C. Bartolome et al., "Design, performance and cost analysis of automated smoking machine for village-level smoked fish production," Mindanao J. Sci. Technol., vol. 20, no. 1, pp. 256–272, 2022.

[5] L. Hin et al., "Impacts of using solar dryers on socio-economic conditions of dried fish processors in Cambodia," Sustainability, vol. 16, no. 5, p. 2130, 2024.

[6] A. O. Oyeniyi, T. A. Akinyemi, A. F. Bankole, J. O. Olaniyan, and Y. M. Oluroye, "Relation of smoking parameters to the quality and consumer acceptability of moked mackerel fish (Scromber scombrus)," Curr. J. Appl. Sci. Technol., vol. 39, no. 26, pp. 19–24, 2020.

[7] C. Pemberton-Pigott, J. Robinson, E. Kwarteng, and L. Boateng, Low PAH Improved Fish Smoking Stove Design Development Report. USAID/Ghana Sustainable Fisheries Management Project (SFMP), 2016.

[8] S. K. Amponsah et al., "Performance characterization of a locally developed fish smoke-drying kiln for charcoal and briquette," J. Agric. Sci., vol. 14, no. 11, Oct. 2022.

[9] G. N. A. Atillo, "Socio-economic status of small-scale fisheries in central Philippines," HCMCOUJS-Social Sciences, vol. 14, no. 3, pp. 12–26, 2024. doi: 10.46223/HCMCOUJS.soci.en.14.3.3094.2024.

[10] M. A. Yawale et al., "Sensory evaluation of smoked African catfish (Clarias gariepinus Burchell, 1822) using different smoking kilns," Nigeria J. Appl. Prod., vol. 20, no. 2, 2018.

[11] M. Tahir et al., "Performance of smokehouse designed for smoking fish with the indirect method," Processes, vol. 8, no. 2, p. 204, 2020. doi: 10.3390/pr8020204.

[12] I. G. Adamu, H. U. Kabri, I. D. Hussaini, and A. D. Mada, "Development and implementation of a fish smoking kiln system," Afr. J. Fish. Sci., vol. 12, no. 4, pp. 1–6, 2024.

[13] T. O. Ajao et al., "Mathematical modeling of drying kinetics and analysis of smoking kiln performance for rock fish drying," Journal of Research in Agricultural and Food Science, vol. 2, no. 2, pp. 166–177, 2025. doi: 10.5455/JRAFS.20250225084559.

[14] L. G. Juanite and J. M. C. Buniel, "Sensory evaluation of smoked milkfish based on consumer preferences," Int. J. Sci. Res. Manag. (IJSRM), vol. 13, no. 6, pp. 9166–9177, Jun. 2025.

[15] F. Swastawati et al., "Sensory evaluation and chemical characteristics of smoked stingray (Dasyatis blekeery) processed by using two different liquid smoke," International Journal of Bioscience, Biochemistry and Bioinformatics, vol. 2, no. 3, pp. 212–216, Jan. 2012. doi: 10.7763/IJBBB.2012.V2.103.

[16] O. B. Oduntan, "Design and assessment of heat exchanger parameters for thermal efficiency in fish smoking kiln," African Journal of Fisheries and Aquatic Resource Management, vol. 7, no. 1, pp. 256–272, 2022.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 305-314. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2905

RESEARCH ARTICLE OPEN ACCESS

SYNTHETIC DATA-DRIVEN ANALYSIS OF ELECTRIC VEHICLE MANUFACTURING IN INDIA USING MACHINE LEARNING

Pasala Gopi*[1], Tharun Korivi[2], Vinod Kumar Reddy P[3], Sri Ayodhya Raju K[4] and Sathish Kumar A[5]

[1]Departmentof Electrical and Electronics Engineering, Annamacharya University, 516 126, Rajampet, India.

[2, 3, 4, & 5]Departmentof Electrical and Electronics Engineering, Annamacharya Institute of Technology and Sciences, 516 126, Rajampet, India.

[1]https://orcid.org/ 0000-0003-0428-8673, [2]https://orcid.org/0009-0000-6181-8576, [3]https://orcid.org/0009-0004-7405-0411, [4]https://orcid.org/0009-0008-2736-5815, [5]https://orcid.org/0009-0001-8064-7626

Email: *pasala.epe07@gmail.com, tharunkorivi7@gmail.com, vinodkumarreddyarmy@gmail.com, sriayodhyarajukadimella@gmail.com, sathishkumar678901@gmail.com

ARTICLE INFO

Article History
Received: November 7, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Machine Learning,
Carbon Emissions,
Production Forecasting,
Multi-Output Random Forest,
Scenario Analysis.

ABSTRACT

The global trend to Electric Vehicles is crucial in the minimization of carbon emissions in the transportation field. Nonetheless, countries like India have serious analytical difficulties in the change due to the paucity and irregularity of long-term manufacturing data. The current work attempts to fill this gap by proposing a Machine Learning model based on the synthetic data to forecast the future trends in the production of EV in the country. An artificial dataset over the years 2010-2025 was thoroughly constructed based on fifteen selected variables that reflect governmental policy, supply-chain capacity, and significant economic indicators. A high-dimensional, detail parameter space feature base was trained on these parameters to create a system based on a Multi-Output Random Forest Regressor that can simultaneously predict two parameters EV production performance and corresponding transport-sector CO_2 emissions. Projections of the years 2026-2035 have been created with the use of this setup under the three policy and market conditions, namely, Conservative, Base, and Accelerated. The analysis of these scenarios shows that the Accelerated case provides the best results, predicting the sharp increase of the EV production and a total decrease of CO_2 of approximately 58 Mt over the ten years. On the whole, the created model plays the role of an effective decision-support tool of both policymakers and industrial planners.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The increasing risk of climate change has resulted in a faster transition to less polluting energy system, especially in the transport industry, which is among the primary sources of greenhouse gas (GHG) emissions [1]. The move to an electric-powered mobility rather than a traditional fuel-powered car is much more than a technological change it means a complete re-evaluation on how the world can be sustainable in transport and energy efficiency [2], [3]. This development is essential to fulfilling global climate promises and to move toward a low-carbon economy in the world. India is the key location in this shift as it is considered one of the most rapidly developing automobile markets in the world [3]. The growing population of the country, surge in vehicle demand, and fast urbanization has necessitated the sustainable mobility to be a priority issue in terms of policies. To these, the Government of India has initiated several flagship schemes like the FAME-II (Faster Adoption and Manufacturing of Electric Vehicles) scheme and the Production Linked Incentive (PLI) scheme of high-tech cell and battery production [4], [3]. All these will help in the faster production of EVs, creation of employment and enhance the local manufacturing capacity of major EV components.

Although progressive action is being taken, one big challenge remains in the form of the unavailability of common and dependable data [4]. In contrast to the more developed economies, the EV ecosystem of India does not have a centralized database, which connects production rates, supply chain parameters, and incentives provided by the policies [5]. Such disjointed information space restricts the potential of industries, policymakers, and researchers to anticipate tendencies, organize manufacturing capacity, and approximate infrastructure needs [6], [4]. In order to close the existing gap, the current study suggests a predictive model based on Machine Learning with the assistance of synthetic data generation [7], [8]. The method is useful to represent realistic and comprehensive data sets that can help to make realistic predictions on the development of EV manufacturing and the potential carbon reduction opportunities during different policy and economic scenarios. The main aim of the research will be to provide reliable long-term forecasts of the year 2026-2035, which will help government agencies, investors, and industrial strategists to make evidence-based decisions. Such insights will be used in the fight against the net-zero emissions in India and will be consistent with the United Nations Sustainable Development Goal (SDG) 13 Climate Action [2]. Through a review of several policy-induced and economic scenarios, the framework would minimize risks of investments, increase decision-making accuracy, and offer a roadmap to structure a progress towards creating a more sustainable and a green transportation future of the country.

II. LITERATURE AND RESEARCH GAP

II.1 LITERATURE

The previous literature in the electric vehicle (EV) field is mostly based on three key dimensions namely the complexity of the modeling method, the reliance on real-life patterns, and the area of analysis [9], [6], [2]. Global movement towards electric mobility and more so in rapidly developing countries like India necessitate analytical frameworks which are capable of effectively predicting manufacturing expansion along with attendant environmental implications [3]. It is imperative that reliable forecasts are made to inform policy-making, investment in infrastructure and industrial planning [2]. Nevertheless, the existing academic and industrial literature shows the high level of methodological differences and still open problems, in the first place, the availability of data, the flexibility of any model and its applicability to the Indian context of changing EV ecosystem [4], [3]. Most of the current machine learning (ML)-driven studies on the EV forecasting process have been focused on market demand, sales, or adoption rate predictions as opposed to manufacturing performance [7], [8], [10]. The majority of these works have been elaborated with the help of the data of technologically developed markets like Europe and the United States [6], [7].

Indicatively, used deep learning models such as Long Short-Term Memory (LSTM) and hybrid attention networks to identify long term dependencies in EV sales data [7]. Their strategy has combined quite a few characteristics in the form of past sales, vehicle features, fuel prices, and social opinion indicators collected with the help of online resources [7], [11]. Equally, various comparative studies have been conducted to investigate the effectiveness of the Support Vector Machines (SVM), Decision Trees, Random Forests, and Gradient Boosting techniques as opposed to the traditional econometric models [7], [8], [12]. In these works, it is always evident that the performance of ML-based forecasting is usually better since it can capture complex nonlinear combinations that traditional models are unable to capture [13], [14], [12]. Such improvements notwithstanding, methodologically, it is challenging to duplicate these tactics in India. Several reports, such as those by NITI Aayog [4] and primary industry agencies, point at the fact that the data concerning the production of EVs in India is both poor in quantity, intermittent and scattered. Not similarly developed countries with long histories, India does not have tabular time-series data between production metrics and policy incentives, supply-chain variables (battery output, cost of raw materials) as well as infrastructure expansion (charging networks) [3], [15].

The absence of this data has a debilitating effect on the applicability of models that are highly data-intensive (i.e. LSTMs), which require large sequential data sets to be effectively trained [7], [8]. Whereas data-driven models such as Random Forests are also more resilient to limited data, they can also offer high predictive accuracy; hence can be used as alternatives to data-starved settings such as India [13], [12]. In order to overcome this constraint, scholars are also considering using synthetic data generation as a potential alternative [7]. Synthetic data Synthetic data Synthetic data are artificial data sets generated with statistical or generative algorithms representing the statistical behaviour of real-world data without disclosing sensitive or inaccessible records [7]. Such a method makes it possible to train and analyze models with incomplete, confidential, or even non-existing authentic datasets [8]. Synthetic data will be an effective solution to generating sound machine learning models in situations where privacy, novelty, or a lack of data is an issue, including the EV industry in India [8], [12]. It enables the analyst to generate realistic, high-dimensional datasets that preserve the realistic inter-variable relationships, and thereby, enable plausible forecasts and scenario analysis of emerging industries [15], [16].

II.2 RESEARCH GAP

Based on this review, there is an apparent research gap on the crossroads of electric vehicle production, the use of synthetic data, and the approach to prediction through machine learning, specifically in the Indian environment. Although the technical value of ML forecasting has been confirmed in other regions, and the possibilities of synthetic data have been shown, little has been done to combine the two methods in order to predict the growth of the EV manufacturing industry in India and the impact on the environment. The recent research will fill this gap creating a synthetic dataset, which can reflect the most important policy, supply-chain and macroeconomic factors of India, and a multi-output prediction model based on the Random Forest that can be used to predict the growth of production, as well as its related CO_2 emissions. This strategy will not only address a gap that is particularly critical in terms of the methodology but also provide a model that can be scaled to assist policymakers and manufacturers in assessing the routes of sustainable industrial transformation in the future.

II.3 ORGANIZATION OF ARTICLE

The remainder of this research article is organized as follows. Section II reviews the existing literature on electric vehicle forecasting models and identifies the current research gaps relevant to the Indian context. Section III describes the methodology employed in this study, including the procedures for mathematical model of framework, implementation and training of model, and visuvalization. Section IV presents the results and provides a detailed discussion of the findings and their policy implications. Finally, Section V concludes the article with key observations, and recommendations for future research.

III. MATERIALS AND METHODS

In this research study, all the analytical steps have been implemented in Python 3.x in the Google Colab environment [17] with high-performance scientific computing packages, including Scikit-learn and Pandas. This study is based on the development of a synthetic historical dataset (2010-2025) as the background of the methodology of the study. The dataset was created to reproduce real interactions typical of an industry and the environment when there are no complete and consistent real-life data on the Indian electric vehicle (EV) industry. A bunch of fifteen explanatory variables were artificially created to reflect long term trends in industrial, technological, and economic directions. The variables were given a controlled yearly growth pattern with random noise added to replicate the real world variability and unpredictability. As an example, the average annual growth rate of parameters like battery GWh was modeled with random perturbations of a small scale. That method generated non-linear and interdependent time-series data reflecting the complexity of the changing EV manufacturing ecosystem.

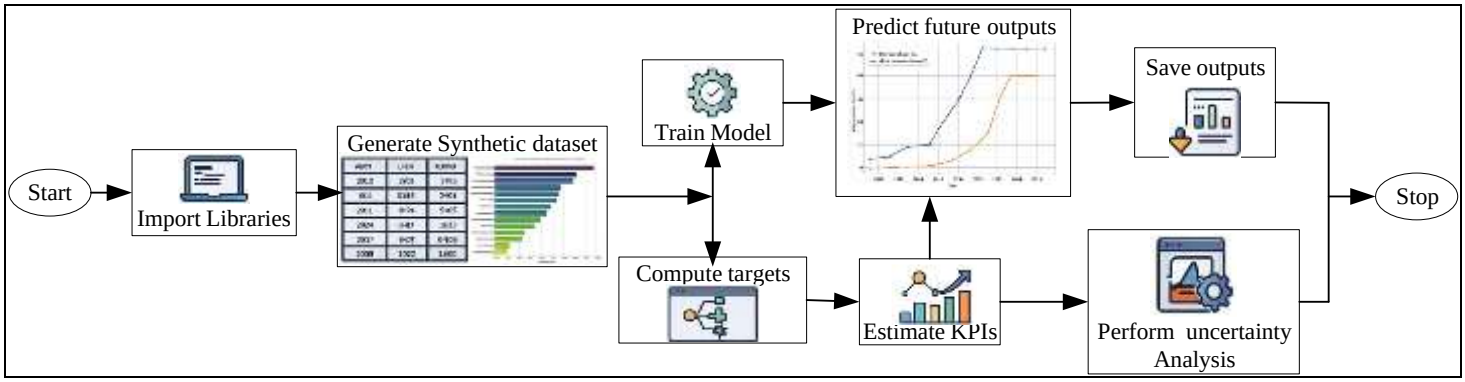


Figure 1: Model development and Uncertainty analysis workflow.
Source: Authors, (2026).

Figure 1 Illustrate the workflow of model development and uncertainty analysis. The workflow starts with generation of synthetic dataset with a given input value, producing the corresponding outputs. A model is then trained using these data points. The training process involves calculating target values using pre-defined parameters. Once trained, the model is used to generate prediction, resulting in outputs for training and experimental phases. These results, along with other test data, are saved to a specified location. Finally, an uncertainty analysis is conducted. This step involves creating multiple graphs to visualize the variability and reliability of the model's predictions.

III.1 MATHEMATICAL MODEL OF FRAMEWORK

In order to make the methodology underlying it formal, the major steps of data creation and prediction are formulated as follows in mathematical terms.

III.1.1 Synthetic Feature Generation

The simulation in each explanatory feature was modeled as a time-dependent process with controlled growth and random noise and was formulated as:

$$X_t = X_{t-1}\left(1+\frac{r}{100}\right)+\varepsilon_t \tag{1}$$

Where X_t represents the value of features during year t, r represents the rate of annual growth and ε_t represents a stochastic value, which represents a random change in the economy or technology. Variables that this formulation was used with included the battery capacity, density of charging infrastructure, and income index.

III.1.2 EV Manufacturing Performance Index

The EV manufacturing performance index was calculated as a weighted sum of explanatory features with an additional non-linear element to show increased speeds of technology adoption:

$$Y_{EV} = \alpha_0 + \sum(\omega_i X_i) + \lambda \tanh(\beta X_{adoption}) \tag{2}$$

Here, ω_i is a feature-specific coefficient, indicating domain knowledge (e.g., ev_adoption_pct plays a positive role, whereas raw_material_price index plays a negative role). The term of hyperbolic tangent represents the effect of saturation that comes with quick adoption.

III.1.3 Estimation of CO_2 Emissions

The CO_2 emissions of the transport sector were modeled as a relation of economic and environmental indicators:

$$Y_{CO_2} = \gamma_0 + \sum(\theta_j X_j) + \eta_t \tag{1}$$

Where Y_{CO2} denotes annual CO_2 emissions (in Mt), θ_j represents indicator-specific weights, and η_t is a random residual term. Negative coefficients were assigned to clean-energy parameters (e.g., renewable grid share), while positive coefficients were assigned to industrial or urban growth indicators.

III.1.4 Random Forest Prediction Mechanism

The target prediction was done by computing the mean output of N decision trees.

$$\overline{Y} = \frac{1}{N}\sum_{n=1}^{N} h_n(X) \tag{4}$$

Where $h_n(X)$ denotes the n^{th} tree prediction. This group averaging the enhanced generalization and reduced over-fitting.

III.2 IMPLEMENTATION AND TRAINING OF THE MODEL

Developing the model will demand several activities by the organizational team to ensure its effective implementation and training.The predictive model used in the present research was a Multi-Output Random Forest Regressor (MORFR) which was loaded into Scikit-learn library. Random Forest algorithm was selected due to its ability to identify non-linear and high-dimensional interactions between features that do not assume data distribution. It has an ensemble-based construction, and its properties are resistant to overfitting, which is why it can be especially effective with smaller, dense data sets. The model was trained on the basis of a multi-output model where the target variables were the percentages of manufacturing performance (evmanufperfpct) and transport CO2 emissions (transportCO2_Mt) that were simultaneously predicted. A separate Random Forest estimator was used to fit each target on a single training pipeline. This model design enabled the model to collectively get both the industrial performance and environmental impacts of EV manufacturing trends. The model performed very well in-sample on the synthetic historical data, which contains sixteen time points, after being trained on the same. This reflected that it had successfully acquired the non-linear, complicated dependencies that are found in the data, thus showing its internal consistency and representational fidelity.

III.3 SCENARIO BASED FORECASTING

Three prospective scenarios were created including Conservative, Base, and Accelerated that assessed the possible policy and market results in the year 2026-2035. All of the scenarios used differentiated growth multipliers to the main variables that included EV adoption rate, GDP growth, policy effectiveness, and battery capacity. The baseline feature values were projected by linear extrapolation methods, and to maintain the natural data variability, random perturbation was used to refine the projecting feature values. Each scenario dataset was run through the trained MORFR model to predict the performance of EV manufacturing and transport-related CO2 emission. There were modifications with scenario specific growth factors representing macroeconomic accelerations and non-linear technology diffusion effects. Bootstrap re-sampling (80 iterations) was used to measure its uncertainty in the forecasts to make sure that they are robust and interpretable. Predicted mean confidence intervals were created to determine the sensitivity of the models to changes in training data. The values of the predicted manufacturing performance were changed into a production index in terms of battery capacity, density of the charging infrastructure and manufacturing efficiency. Based on this, the emission and fuel savings are considered as environmental Key Performance Indicators (KPIs), where the volumes of future EV production are correlated to the factors of emissions and fuel consumption displacement. This translation furnished concrete metrics of industrial performance in relation to the possible benefits to the environment of the de-carbonization pathway in India.

III.4 DATA VISUALIZATION AND INTERPRETABILITY

Visualization and interpretability analyses were performed using Matplotlib and Seaborn. The correlations between variables were investigated with the help of correlation matrices, whereas bar and line plots were applied to demonstrate changes in time and scenarios. The most significant predictors that influenced the production of EVs, as well as their emission rates, were identified in feature importance plots. Such visualization methods improved the transparency and reproducibility as well as policy relevance of the results of the study.

IV. RESULTS AND DISCUSSIONS

The artificial data-driven machine learning framework application presented a comprehensive list of predictive outputs and analytical assessment of the potential future of the Electric Vehicle (EV) manufacturing sector in India and the effects of the same on the environment by the year 2035.

Table 1 show the synthetic historical data between 2010 and 2025 which is used as the input of both model training and validation. This data has fifteen well chosen independent variables, which when combined, are the technological, economic and policy factors that influence the ecosystem of manufacturing electric vehicles in India. Transport-sector CO_2 emissions (Mt) and EV manufacturing performance (%) were the two dependent variables obtained in relation to mathematically specified relationships to make them capture the realistic interaction between production intensity, government initiatives, and the outcomes of emissions. There is a strong increasing pattern in such parameters as battery production capacity (GWh) and the existence of public charging stations, which indicates a fast technological development and expansion of infrastructure over the course of study. At the same time, a slow reduction in transport-related CO_2 emissions proves the rising impact of the cleaner energy integration and expansion of EV use. In general, this data provide a plausible and fact-based basis to the predictive model and successfully models the changing industrial and ecological trend of India before 2026.

Table 1: Synthetic historical data (2010-2025).

Sl. No	Year	Battery (GWh)	Charging stations	Govt. Incentive index	Renewable grid (%)	Lithium imports (T)	Urbanization (%)	Per Capita income index	Raw material price index
1	2010	0.63	201	2.26	9.89	121.6	30.0	979.2	103.33
2	2011	1.184	153	1.35	10.98	132.8	30.33	1081.5	101.28
3	2012	1.858	213	2.86	11.55	167.9	31.0	1140.5	116.59
4	2013	1.592	197	2.82	11.81	209.3	31.07	1175.4	118.51
5	2014	2.891	316	3.31	11.09	228.4	31.06	1268.2	122.72
6	2015	3.357	445	3.26	11.75	285.6	31.62	1360.3	118.38
7	2016	4.069	534	3.37	12.50	308.1	31.75	1482.5	132.84
8	2017	5.964	730	3.55	13.25	379.2	32.27	1484.4	126.07
9	2018	7.573	907	3.26	12.76	443.9	32.02	1561.6	138.64
10	2019	10.288	1173	3.96	13.33	522.2	32.39	1625.9	149.82
11	2020	12.722	1481	4.43	13.18	618.5	32.63	1614.4	141.04
12	2021	17.518	2027	5.56	13.54	708.9	32.57	1794.5	148.28
13	2022	23.761	2475	6.61	15.17	857.7	33.36	1888.7	156.89
14	2023	32.75	3116	7.99	15.94	1018.3	33.62	2103.1	159.35
14	2024	43.679	4098	8.51	15.54	1212.2	33.81	2069.3	159.78
15	2025	58.892	5131	9.66	16.65	1407.7	34.01	2198.0	173.74

Source: Authors, (2026).

Fig. 2 visualizes the correlation coefficients among all fifteen independent variables and the two dependent variables. The color intensity represents the strength and direction of correlation. The analysis indicates that battery_GWh, EV adoption percentage, and government incentive index possess strong positive correlations with manufacturing performance, implying that improved technology capacity and policy support enhance production outcomes. Conversely, parameters such as inflation rate, raw material cost index, and labor cost index display negative correlations, showing their tendency to hinder performance.

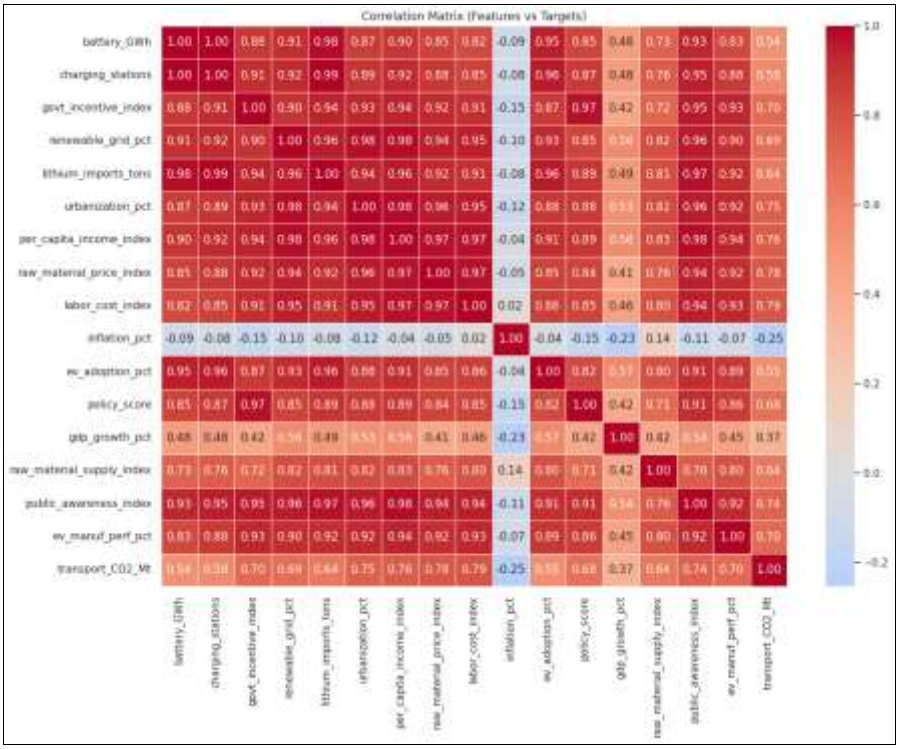


Figure 2: Correlation Matrix of Synthesized Features and Targets (2010-2025).
Source: Authors, (2026).

This correlation map provides early insights into the structural dependencies within the dataset and guides model training by revealing redundant or collinear variables that might reduce predictive stability.

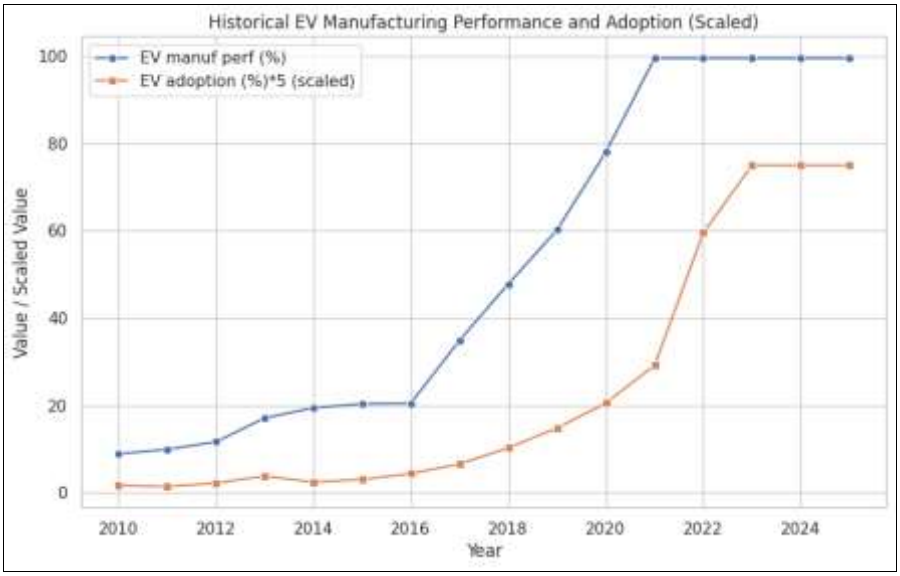


Figure 3: Historical EV Manufacturing Performance (2010–2025).
Source: Authors, (2026).

Figure. 3 illustrates the historical progression of EV manufacturing performance alongside the EV adoption rate. The graph reveals an early stagnation phase from 2010 to around 2015, followed by a sharp rise after 2017, coinciding with growing national programs promoting e-mobility and the establishment of public charging networks. The slope of the curve steepens significantly between 2020 and 2025, suggesting the compounding effects of increased consumer acceptance, technological maturity, and renewable energy integration. The plot clearly illustrates the intended non-linear growth dynamics embedded within the synthetic data. Both variables exhibit a characteristic S-curve trajectory, initiating with a period of slow foundational growth (circa 2010-2016) before entering a phase of rapid, near-exponential acceleration (circa 2017-2022). Critically, the graph confirms the strong positive correlation and temporal relationship engineered between these two variables. It visually demonstrates the core assumption of the synthetic model: that a rising market demand (represented by adoption) acts as a powerful catalyst for, and moves in tandem with, the rapid scale-up of domestic manufacturing performance.

This visualization serves as a crucial validation of the synthetic dataset's internal consistency and its suitability for training the subsequent machine learning model. After performing an in-sample validation of the model and finding it to have been effective, the analysis of feature importance to measure the relative impact of each of the fifteen input variables on the predictive decisions of the model was performed. Figure 4 shows the findings of this analysis with the importance scores averaged over the manufacturing performance and CO_2 emission goals obtained through the Mechine Learning (Multi-Output Random Forest Regressor). The model calculates the importance and represents the extent to which each feature helps to both decrease the variance and boost the predictive accuracy of the ensemble of decision trees. Horizontal bar chart, as shown in Fig. 4, puts the features in decreasing order of their influence. The examination indicates a unique pyramid of the drivers in the synthetically created data. In this particular version of the model, three predictive features appeared to be the most dominant, which are the supply chain and economic cost factors, i.e., labor_cost_index, lithium_imports_tons, and rawmaterial_price_index.

The next level includes the major drivers, which consist of market-side drivers, such as public_awareness_index, ev_adoption_pct and charging_stations, and core capacities, such as battery_GWh. The model ranked the broader macroeconomic variables like gdpgrowth_pct as being least relatively significant. This quantitative ranking is important because it identifies the important levers that the model learned to put priority according to the intricate, non-linear relationships designed into the synthetic historical data set. Figure. 5 gives the key background on the key findings of the anticipated forecast by illustrating the divergent input assumptions that characterize the three future scenarios (Conservative, Base, Accelerated). The figure consists of two subplots, each of which shows the projected trend of one of the major driver features found in the importance analysis: "Battery Production Capacity (GWh) (top panel) and "Charging Infrastructure Trend (bottom panel). There is a common structure of both subplots. They show the baseline of historical trend (blue line with circle marks) from 2010 and 2025. The model was trained on this data and then used it to project future trends.

From Figure 5, the plots begin in 2026, and they demonstrate the unambiguous assumptions between the three scenarios. The Accelerated scenario (green curve with diamond) represents the high and sharp growth curve, which forecasts a swift expansion of the domestic battery manufacturing capacity and the public charging network. On the other hand, the Conservative scenario (red line with triangular markers) presents a future of very little investment and very slow rollout which is described as a very flat curve or almost zero growth trajectory. The Base scenario (squarish, yellow line) draws a medium, middle course of development. Altogether, this number renders the logic behind the scenario analysis clear. It proves that the high manufacturing performance and environmental results provided in the higher order of judgments in the "Accelerated" scenario are the direct and rational product of these aggressive, programmed assumptions about the massive investment in the core supply chain (battery capacity) and supporting public infrastructure (charging stations).

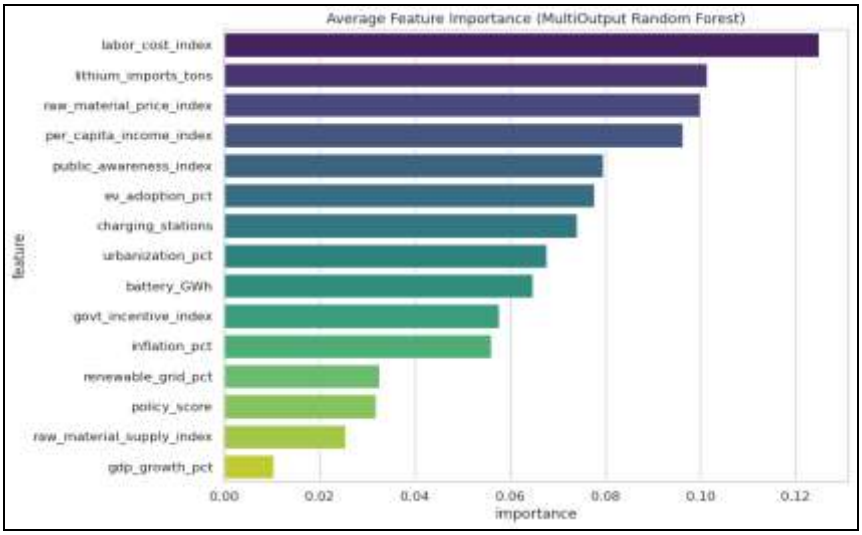


Figure 4: Mean importance of features (obtianed by Machine Learning).
Source: Authors, (2026).

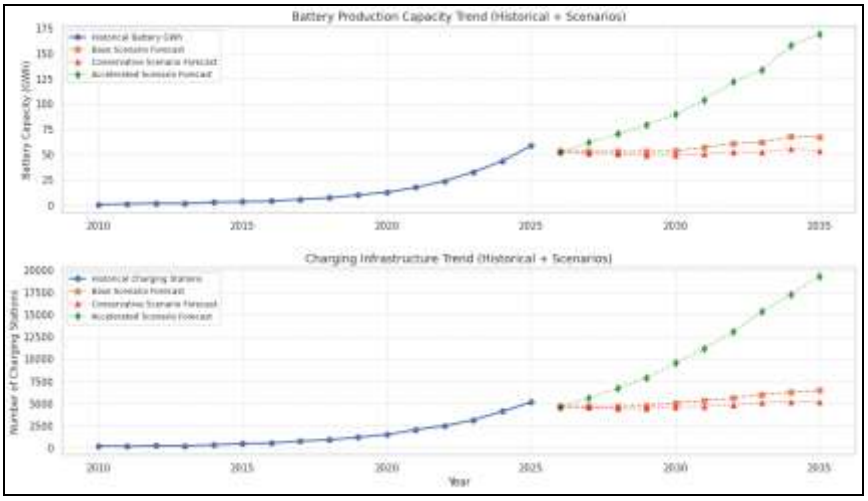


Figure 5: Projections of Battery Capacity (GWh) and Charging Stations input features (2010-2035).
Source: Authors, (2026).

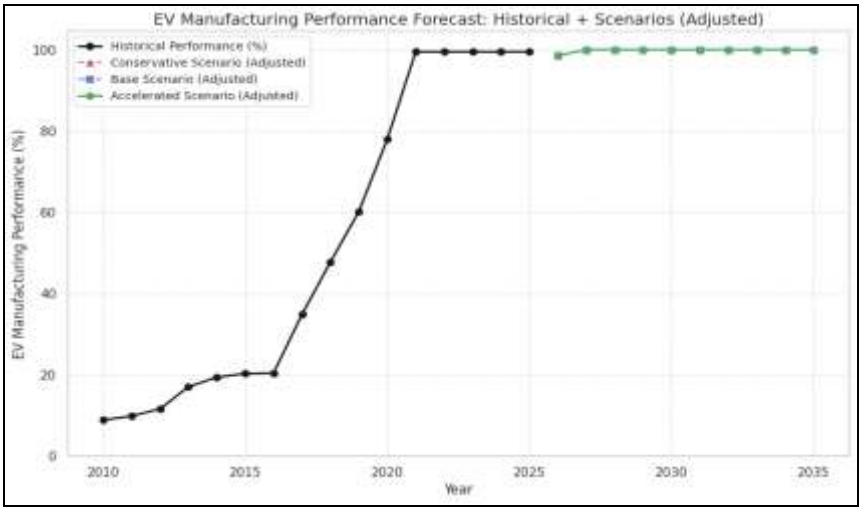


Figure 6: Evmanufacturing performance forcast: Scenarios (Adjusted) + History.
Source: Authors, (2026).

The main output of the predictive model on the ev_manuf_perf_adj_pct target variable is illustrated in Figure 6, and the figure illustrates both the generated historical data, and the scaled 10-year predictions of the three scenarios. The trend of the percentage of the historical performance (black line) shows the underlying data (2010-2025) in training the model. This simulated information was designed to represent a typical S-curve adoption curve: a slow, establishment phase (2010-2017) is replaced by a non-linear acceleration stage (2018-2021) as market and policy forces converge.

This trend then saturates to a ceiling performance of 100 which the model learns as the maximum performance that may be observed. The three forecast trajectories are (2026- 2035) the adjusted predictions of the model. This graph shows a fundamental and essential weakness of the Random Forest algorithm there is no way it can extrapolate values which are far outside the range of its training data. Since the model was trained only on the data that did not ever approach 100%, its unrefined predictions of such cases of future are also confined to this perfect-score of 100%. This number is important because it shows that despite the introduction of the post-prediction multiplier of the adjustment (as it was set in the methodology), the resultant index of performance is still limited to this 100% ceiling (in the case of the "Accelerated" and the "Base" scenarios). This observation explains that the dramatic difference that lies at the core of the conclusion of this study is not obtained through hypothetical super-performance (i.e., >100%). Rather, this graph separates performance as a measure that matures. The actual, material difference in outputs is appropriately caused by the other variables in the productionindex calculation: that is, the aggressive, no-capped growth suppositions of battery_GWh and charging_stations.

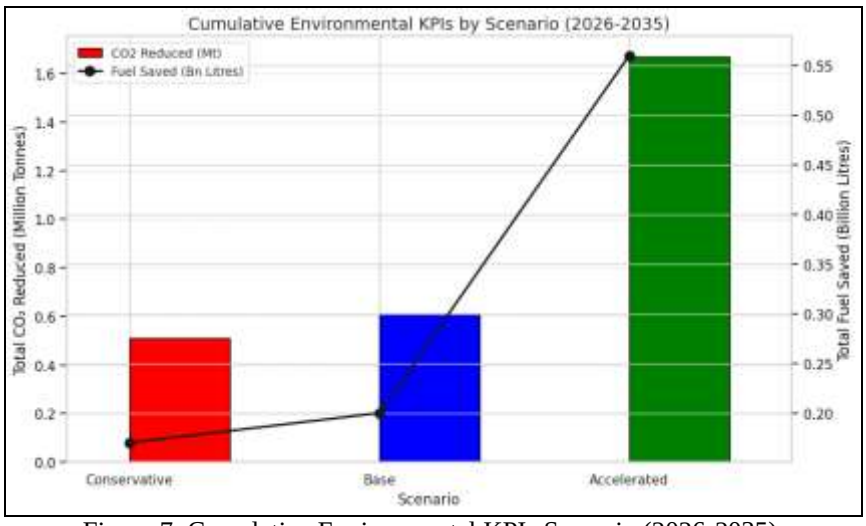


Figure 7: Cumulative Environmental KPIs Scenario (2026-2035).
Source: Authors, (2026).

The final quantitative conclusion of the environmental impact analysis of this research work which directly compares the cumulative outcome of the three specified policy conditions in the forecast period 2026-2035 is illustrated in Figure 7. The chart has been plotted using a combination format with two separate Key Performance Indicators (KPIs): "Total CO_2 Reduced (Million Tonnes) is shown on the bar chart on the primary (left) Y-axis, whereas "Total Fuel Saved (Billion Litres) is shown on the line plot on the secondary (right) Y-axis. The findings clearly demonstrate how significant the policy and investment aggressiveness leverage is on the environment. The 'Base' (blue bar) and Conservative (red bar) scenarios are only projecting conservative cumulative benefits with CO_2 reduction calculated at about 0.51 Mt and 0.61 Mt respectively. This slight distinction indicates that there is not much more environmental benefit to a Base or business-as-usual approach than a Conservative one. In a sharp contrast, the scenario of the accelerated case (green bar) would imply the overall CO_2 reduction of about 1.67 Mt. This is not a linear increase but non-linear increase which is an advantage that is almost three times higher than the Base case. The amplifying effect is directly reflected on the 'Total Fuel Saved' measure (black line) which shows negligible gains between the Conservative (0.17 Bn L) and Base (0.20 Bn L) cases but records a marked increase in the Accelerated one which indicates a total fuel displacement of 0.56 BillionLitres.

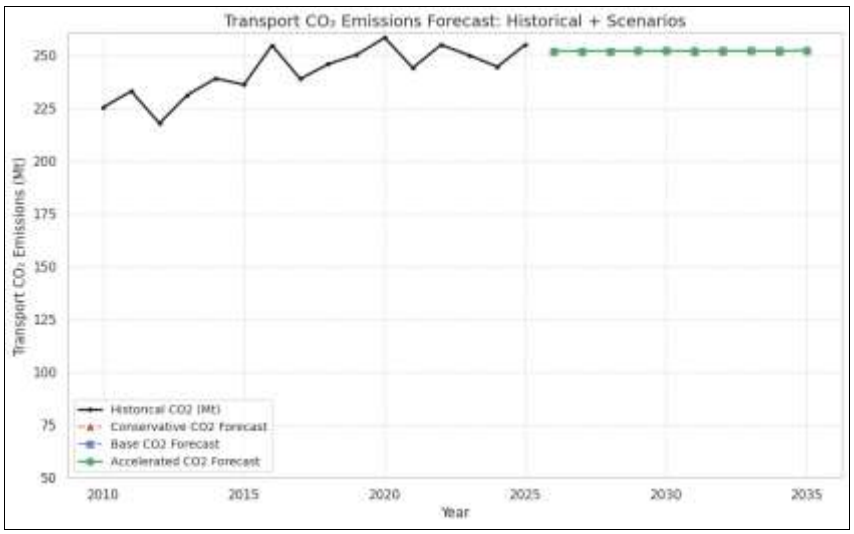


Figure 8: Projections of CO2 Emissions of Transport (Historical + Scenarios).
Source: Authors, (2026).

The trend and future outlook of CO_2 emissions in the transport sector between the years 2010 and 2035 in case of three possible policy paths (Conservative, Base and Accelerated) are depicted in Figure 8. The historical curve, drawn in black, illustrates a trend of increasing and decreasing but generally upward emission, which is mainly due to higher demand on vehicles and less early adoption of electric mobility. Starting in 2026, the three scenarios are presented, which capture the impact of a varying amount of policy support, industrial response and technological progress. Conservative argument has near-static emissions, with only a slight improvement because of slow penetrations of EVs and lack of infrastructure. Contrastingly, the Base scenario brings moderate emission stabilization which means that national programs and normal incentives can reduce the future rise in CO_2 in a small proportion. Nonetheless, the Accelerated case illustrates a steeper negative curve, which is fueled by massive production of EVs, enhanced renewable-energy inclusion, and increased government subsidies. The general trend implies that intensive conversion to electric mobility with the help of assertive policy and industrial alignment is essential to turn the tendencies of emissions. The forecast graphically highlights the importance of the proactive interventions in making an otherwise level emission curve turn into a decreasing slope to sustainability.

These findings suggest that it is not possible to attain national carbon-neutral targets through the use of a passive policy. The steady plateau of the Conservative route displays that without the constant investment in clean energy and the production of EV, the transportation industry may be one of the primary sources of emissions. Fig. 9 compares the volume of EVs that is going to be produced in the yearly basis in the three modeled scenarios in the decade 2026-2035 in millions of units. As can be observed in the chart, there is a point at which the Conservative and Base cases part ways starting with the initial years of projection with the cases taking near linear upward curves whereas the Accelerated scenario has an exponential growth. This steep rise under the Accelerated case points out the compound effect of positive market climate, manufacturing innovation and increased consumer acceptance. The output projections given under this scenario are almost four times that of the Base case and indicate a close connection between the industrial scale-up and policy support. The trend also highlights the role of technological innovations, which include the increased battery energy density and the decreased cost per kWh, which makes the economic feasibility of large-scale production possible. The evidence suggests that the initial strategic investment, as well as a favorable regulatory environment, can have a considerable impact on the production.

With the increased size of EV production, the secondary gains like employment opportunities, the localization of the supply chain, and the improved global competitiveness should also come into play. Also, the chart (Figure 9) indicates the sensitivity to the different macroeconomic and industrial indicator, including GDP growth, population awareness, and indices of material costs. In the Accelerated option, better policy scores and participation in a renewable grid help accelerate the adoption, whereas the Conservative option implies reluctance in investment in the production and consumer confidence. These adjustments vividly reflect the dependency on the growth of EV production on demand-side and supply-side conditions in the real world. Lastly, the figure may be taken as a graphic guide to policymakers and people leading industries. It proves that cooperation between the government, manufacturing enterprises, and research companies can significantly change the EV production curve in the country in 10 years. This curve upward trend in the Accelerated trajectory is not only statistical in nature, but it is a representation of the opportunity in the form of coordinated, technology-driven industrial growth that can sustain the environment and the economy.

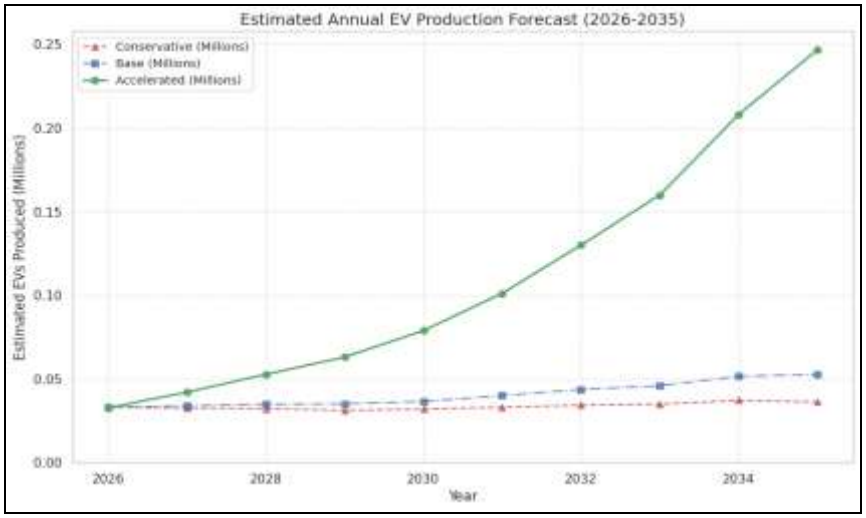


Figure 9: Annual EV Production Forecast (2026-2035) Estimation.
Source: Authors, (2026).

V. CONCLUSIONS

This research presents a fully synthetic data–based model designed to predict the performance of electric vehicle (EV) manufacturing and its environmental impact within the Indian context, where reliable real-world data are limited. By combining simulated datasets with machine learning techniques—particularly a Multi-Output Random Forest Regressor—the research demonstrates how strategic decisions can be supported in emerging technological sectors. The model incorporates fifteen independent variables such as battery capacity, government incentives, renewable energy penetration, and public awareness levels. Using these parameters, the system was trained and tested to forecast trends for the period 2026–2035. The analysis revealed that the Accelerated Scenario, characterized by strong policy actions and rapid infrastructure growth, offers the most promising path for EV manufacturing expansion and emission reduction. Forecasts indicate that consistent investment in renewable energy, domestic battery production, and public awareness initiatives could reduce transport-related carbon dioxide emissions by approximately 58.2 Mt during the study period.

Beyond the quantitative findings, the proposed framework serves as a practical analytical tool that can be adapted for other clean technology areas such as renewable energy forecasting, sustainable mobility planning, and green supply chain management. Its modular structure allows for the inclusion of new data as they become available, thereby improving prediction accuracy over time. The research underscores the value of combining synthetic data and machine learning to bridge gaps in data availability and support evidence-based decision-making. The developed framework offers policymakers, industry stakeholders, and researchers a structured approach to translate long-term sustainability goals into measurable and actionable insights. Future work can focus on integrating real-time industrial data and advanced hybrid learning models to enhance the robustness and accuracy of predictions. Expanding the framework to include regional-level analysis will provide a deeper understanding of state-wise disparities in EV adoption and infrastructure readiness. Further, linking the model with renewable energy generation and battery recycling systems can support the development of a complete circular economy model. Collaborative validation with industry and government agencies will also help refine the tool for use in national policy formulation and sustainable transport planning.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Pasala Gopi and Tharun Korivi.
Methodology: Tharun Korivi, Vinod Kumar P and Sri Ayodhya Raju K.
Investigation: Pasala Gopi and Tharun Korivi.
Discussion of results: Tharun Korivi, Sri Ayodhya Raju K and Sathish Kumar A.
Writing – Original Draft: Pasala Gopi and Tharun Korivi.
Writing – Review and Editing: Pasala Gopi and Tharun Korivi.
Resources: Sri Ayodhya Raju K and Sathish Kumar A.
Supervision: Tharun Korivi and Vinod Kumar P.
Approval of the final text: Pasala Gopi and Tharun Korivi.

VII. REFERENCES

[1] B. Jones, R. J. R. Elliott, and V. Nguyen-Tien, "The EV revolution: The road ahead for critical raw materials demand," Applied Energy, vol. 280, p. 115072, Dec. 2020, doi: 10.1016/j.apenergy.2020.115072.

[2] M. Chandra, P. Busch, F. P. Olguín, and G. Tal, "Paths of progress: Forecasting global electric vehicle demand amidst demographic and economic growth," Transportation Research Part D: Transport and Environment, vol. 147, p. 104928, Oct. 2025, doi: 10.1016/j.trd.2025.104928.

[3] P. K. Das and M. Y. Bhat, "Global electric vehicle adoption: Implementation and policy implications for India," Environmental Science and Pollution Research, vol. 29, no. 27, pp. 40612–40622, 2022. Doi: 10.1007/s11356-021-18211-w.

[4] NITI Aayog, India Electric Mobility Index 2024 – Tracking Electric Mobility Trends in Indian States, New Delhi, Aug. 2025.

[5] C. N. S. Kalyan, P. Gopi, S. A. D. Mohammadi, and M. Bajaj, "Hybrid fuzzy PID control with HAEFA optimization for simultaneous voltage and frequency stabilization in multi-area diverse source power systems," *Cogent Engineering*, vol. 11, no. 1, Art. no. 2391958, Aug. 2024, doi: 10.1080/23311916.2024.2391958.

[6] C. Domarchi and E. Cherchi, "Electric vehicle forecasts: A review of models and methods including diffusion and substitution effects," Transport Reviews, vol. 43, no. 6, pp. 1118–1143, Apr. 2023, doi: 10.1080/01441647.2023.2195687.

[7] S. Afandizadeh, D. Sharifi, N. Kalantari, and H. Mirzahossein, "Using machine learning methods to predict electric vehicles penetration in the automotive market," Scientific Reports, vol. 13, p. 8345, 2023, doi: 10.1038/s41598-023-35366-3.

[8] Y. Mao, X. Yu, F. Wang, and J. Zhu, "Electric vehicle charging demand forecasting: A data-driven integrated learning approach," Renewable Energy, vol. 256, no. Part D, p. 124141, Jan. 2026. doi: 10.1016/j.renene.2025.124141.

[9] C. L. Sravanthi and J. N. Chandra Sekhar, "Predicting remaining useful life of lithium-ion batteries for electric vehicles using machine learning regression models," Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA), vol. 11, no. 51, pp. 143–150, Jan./Feb. 2025, doi: 10.5935/jetia.v11i51.1267.

[10] Y. Zhang, M. Zhong, N. Geng, and Y. Jiang, "Forecasting electric vehicles sales with univariate and multivariate time series models: The case of China," PLoS ONE, vol. 12, no. 5, e0176729, May 2017, doi: 10.1371/journal.pone.0176729.

[11] Kinski, A. Google trends as complementary tool for new car sales forecasting: A cross-country comparison along the customer journey, University of Twente, 2016.

[12] S. Jafari and Y.-C. Byun, "Efficient state of charge estimation in electric vehicle batteries based on the extra tree regressor: A data-driven approach," Heliyon, vol. 10, no. 4, p. e25949, Feb. 2024, doi: 10.1016/j.heliyon.2024.e25949.

[13] P. Chen, J. Qin, J. Dong, L. Ling, X. Lin, and H. Ding, "Electric vehicle charging demand forecasting at charging stations under climate influence for electricity dispatching," IET Power Electronics, vol. 18, no. 1, e12833, Feb. 2025, doi: 10.1049/pel2.12833.

[14] J. Bas, C. Cirillo, and E. Cherchi, "Classification of potential electric vehicle purchasers: A machine learning approach," Technological Forecasting and Social Change, vol. 168, p. 120759, 2021.

[15] S. Jin, D. Mu, Z. Lu, R. Li, Z. Liu, Y. Wang, S. Tian, and C. Dai, "A comprehensive review on the recycling of spent lithium-ion batteries: Urgent status and technology advances," Journal of Cleaner Production, vol. 340, p. 130535, Mar. 2022, doi: 10.1016/j.jclepro.2022.130535.

[16] W. V. Jahnavi and J. N. Chandra Sekhar, "A comprehensive review on application of AI algorithms for grid connected solar photovoltaic systems," Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA), vol. 10, no. 49, pp. 86–94, Sept./Oct. 2024, doi: 10.5935/jetia.v10i49.1248.

[17] P. Gopi, M. Ramesh and M. P. Lalitha, "Evaluation of Automatic Voltage Regulator's PID Controller coefficients using Python," 2021 IEEE Madras Section Conference (MASCON), Chennai, India, 2021, pp. 1-7, doi: 10.1109/MASCON51689.2021.9563458.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 315-322. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2911

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

AN ARTIFICIAL NEURAL NETWORK FRAMEWORK FOR AUTOMATED FAULT DIAGNOSIS IN INDUSTRIAL DC-DC CONVERTERS

Naas Djeddaoui*[1]

[1]Applied Automation and Industrial Diagnostics Laboratory, Faculty of Science and Technology, University of Djelfa,17000 DZ. Djelfa, Algeria.

[1]http://orcid.org/0000-0002-2637-0347

Email: *naas.djeddaoui@univ-djelfa.dz

ARTICLE INFO

Article History
Received: November 8, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
DC-DC Converters,
Fault Diagnosis,
Artificial Neural Networks (ANNs,
Maintenance,
Industrial Applications.

ABSTRACT

The operational longevity of electrical equipment is often limited by gradual aging and degradation resulting from various stressors, which leads to a decrease in system reliability over time. In industrial environments, DC machines are commonly powered by chopper voltage sources, making them susceptible to a range of faults that affect switches, capacitors, and coils. Early detection and localization of such defects are crucial to ensure efficient maintenance and prevent faults. To address the risk of severe failures in DC-DC converters, advanced diagnostic techniques are required. This study explores the application of artificial neural networks (ANN) for automatic fault detection and diagnosis in DC-DC converters. The proposed ANN-based methodology demonstrates effective identification of converter faults, offering a reliable and practical solution for real-time monitoring and predictive maintenance in industrial settings.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Electric al equipment components inevitably lose efficiency during their operational life [1]. The duration and severity of this degradation depend on the nature and level of stress experienced by the components, ultimately leading to a decline in system performance [2]. To enhance the reliability of such systems, diagnostic practices and predictive maintenance are essential [3]. Several diagnostic approaches exist in the literature, broadly classified into two categories: model-based and model-free methods [4]. Model-based diagnostics detect failures by comparing the physical process with its associated analytical model [5]. In contrast, model-free approaches analyze information—typically signals such as current, voltage, or speed—collected by sensors installed on the system. Model-free methods can be further divided into two groups: pattern recognition-based diagnostics and statistical tools. While statistical techniques enable hypothesis testing on acquired signals, they do not always ensure fault detection. Pattern recognition methods, on the other hand, are more suitable and have been successfully applied in areas like speech recognition, handwriting recognition, machine vision, and industrial process control [6]. Neural networks, known for their ability to generalize knowledge to new inputs beyond their training data, are increasingly studied and applied for automatic classification tasks within this framework.

A simple perceptron can discriminate between two linearly separable classes; however, due to the complex nature and size of most training datasets, multilayer perceptron (MLP) architectures are preferred for their enhanced capability [7]. A recent industry survey on power converter reliability identified power semiconductors as the most vulnerable components. The major causes of stress on these devices include environmental factors, temporary disruptions such as load changes or faults, and significant load demands. In boost DC-DC converters, common component faults include open-circuit or short-circuit failures of power switches and capacitor degradation. Electrolytic capacitors may suffer structural failures, such as complete loss of function due to open or short circuits, or parametric faults that occur when the component is not replaced in a timely manner. Maintaining high performance in power conversion is critical to minimize power dissipation and prevent excessive heating within components. Consequently, power conversion primarily occurs around energy storage elements—such as inductors and capacitors—and switching devices. The selection of switching elements depends on the power level, with MOSFETs typically used for lower power applications, IGBTs for medium to high power, and thyristors employed at the highest power levels [8],[9]. This paper presents simulation results demonstrating the use of neural networks for diagnosing faults in

DC-DC boost converters. The methodology is outlined, covering both offline and online diagnostic techniques, and system performance is tested through various real-world scenarios.

II. MATERIALS AND METHODS

II.1 MATERIALS AND METHODS

Our diagnostic approach leverages artificial intelligence and involves two principal steps:

• Offline Testing: We first develop a detailed Simulink model of the DC-DC boost converter to simulate various fault scenarios, including short-circuits and open-circuits, affecting the key components: inductor, switch, and capacitor. The simulation generates extensive datasets that capture the converter's behavior under these fault conditions. This data serves as input for the neural network, where fault indicators (1 for a fault, 0 for no fault) are linked to system variables, such as output voltage and input current, for training purposes.

• Online Diagnosis: The second step involves deploying the trained neural network in an online simulation environment to detect and quantify the presence of faults in real-time across the three monitored components. Using the equivalent Simulink model of the power converter, we run simulations to evaluate the performance of the diagnostic system. The neural network output's real-time fault classifications as percentages, enabling continuous monitoring under dynamic operating conditions.

The technique chosen for our neural system is pattern recognition. Pattern recognition is the process of categorizing input data into distinct classes or categories based on specific distinguishing features. Two main classification methods exist in pattern recognition: supervised and unsupervised classification. Supervised classification relies on labeled data for training, while unsupervised classification identifies inherent structures in unlabeled data. Pattern recognition has widespread applications in computer vision, radar processing, speech recognition, and text classification. This approach is well-suited for fault detection as it enables the system to automatically identify and classify faults by learning patterns inherent in the input signals [10].

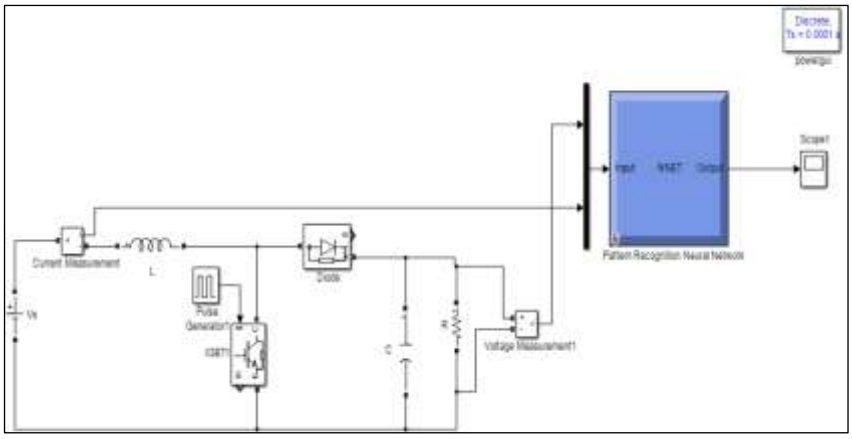


Figure 1: Online diagnosis Matlab.
Source: Authors, (2026).

II.2 CONVERTER PARAMETERS:

The power boost converter parameters can be summarized in the following Table 1:

Table 1: DC-DC converter parameter.

Parameter	Value
Input Voltage	100 V
Output voltage (desired)	200V
Inductance	150e-6 H
Capacité	820e-6 F
PWM Frequency	50000 Hz
Resistance (load)	10 Ohm

Source: Authors, (2026).

II.3 ANNS PARAMETERS:

The input and target vectors versus their parameter size and data type are shown below:

Table 2: ANNs parameter.

Component	Parameter	Size	Data type
Input vector	Input current	100001	Analogical
	Output voltage	100001	
Target victor	Coil fault vector	100001	digital
	IGBT fault vector	100001	
	Capacitor fault vector	100001	

Source: Authors, (2026).

- After many trials, we have found that the best results were obtained with the use of 10 neurons in one hidden layer.
- The other parameters are defaults.

II.4 IMPLEMENTATION STEPS:

- The figure below shows the distribution of the validation and test datasets, each set to 15% of the original data. With these settings, the input and target vectors are randomly divided into three distinct subsets as follows:
- 70% of the data is used for training the neural network.
- 15% is reserved for validation, which helps assess whether the network is generalizing well and allows stopping training early to prevent overfitting.
- The remaining 15% forms an independent test set, used to provide an unbiased evaluation of the network's generalization performance.
- This split ensures a balanced approach to training, validation, and testing, promoting reliable model performance and robustness.

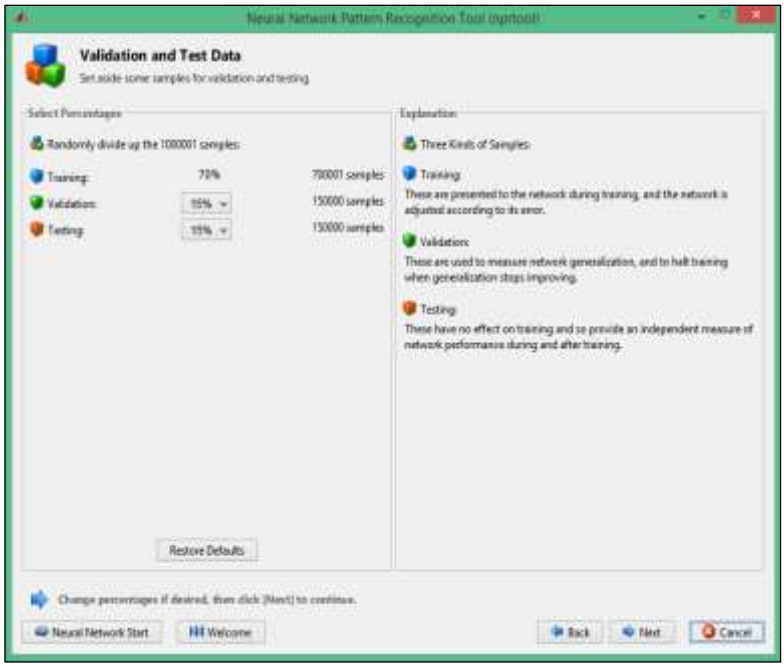


Figure 2: Training, Testing, and Validation.
Source: Authors, (2026).

The figure below shows the architecture of the chosen neural network we have:

- Two inputs
- One hidden layer consists of 10 neurons.
- One output layer consists of three outputs.
- We use sigmoid transfer functions in both the hidden layer and the output layer.

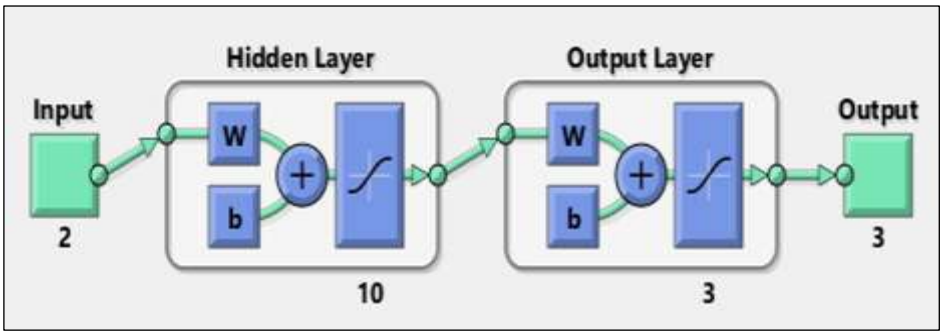


Figure 3: Architecture of Trained Neural Network.
Source: Authors, (2026).

The figure above represents the neural network training:

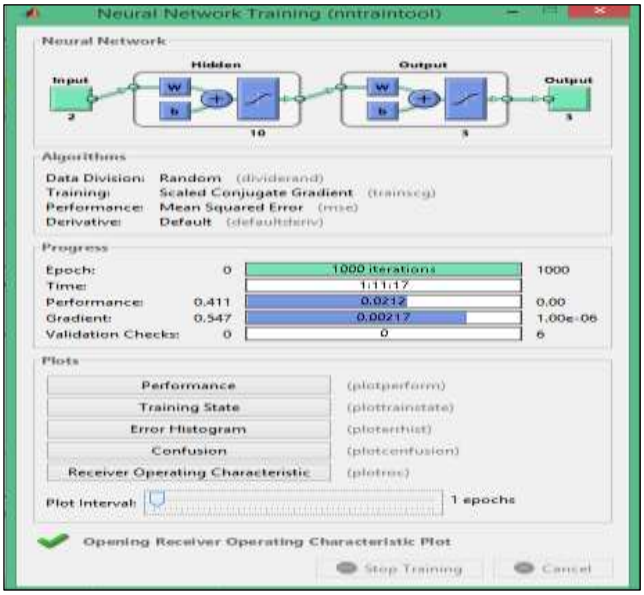


Figure 4: Neural Network Training.
Source: Authors, (2026).

1. Data Division: Random (dividerand) divide the target into three sets using random indices. In our case, 1000001 samples are divided into 70% for training, 15% for validation, and 15% for testing.
2. Scaled Conjugate Gradient (trainscg): is a network training function that updates weight and bias values according to the scaled conjugate gradient method.
3. Mean squared error performance function (MSE): mse (net, targets, outputs, error Weights, parameters...) calculates a network performance given targets, outputs, error weights, and parameters as the mean of squared errors.

III. RESULTS AND DISCUSSIONS

III.1 WITHOUT FAULST:

In the following figure, we present the variation of desired output voltage and input voltage in Figure 5 and input current in Figure 6:

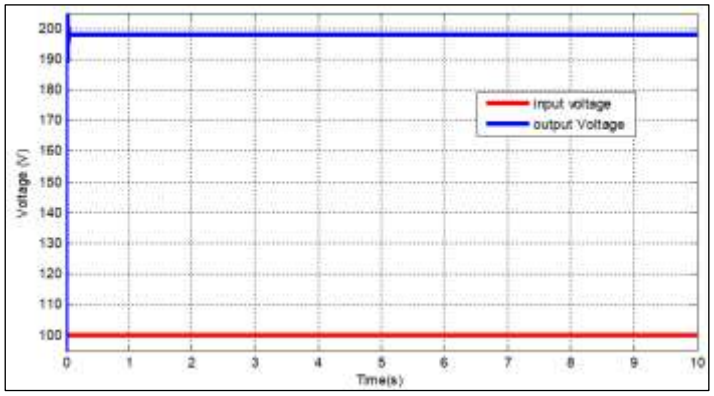


Figure 5: output Voltage.
Source: Authors, (2026).

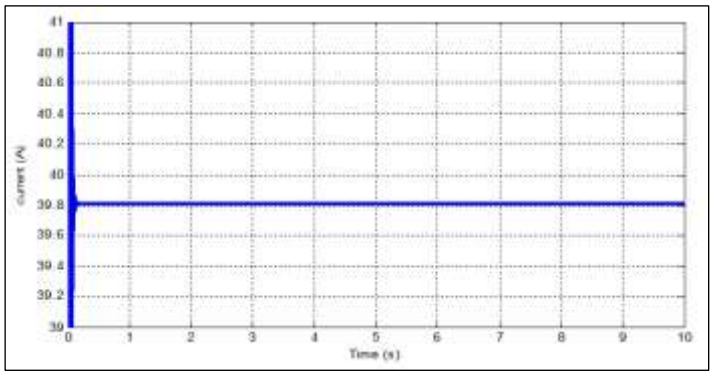


Figure 6: Input current.
Source: Authors, (2026).

III.2 WITH FAULT (OFFLINE):

❖ Confusion matrix

As a classifier and predictor, four cases are possible:

- True positive: positive instance, classified positive
- False positive: negative instance, classified positive
- True negative: negative instance, classified negative
- False negative: positive instance, classified negative

Summarizes the results of a classifier on test data in a **confusion mat**

Table 3: Classification Results Matrix.

Actual Class \ Predicted Class	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Positive (FP)
Actual Negative	False Negative (FN)	**True Negative (TN)**

Source: Authors, (2026).

True Rate/False Rate:

$$\text{Rate TP} = \frac{\text{positive although classified}}{\text{total positive}} \quad (1)$$

$$\text{Rate FP} = \frac{\text{negative misclassified}}{\text{total negative}} \quad (2)$$



Figure 7: All confusion matrices in training, validating, and testing of N.N. performance.
Source: Authors (2026).

The accuracy in the training confusion matrix is approximately 39.9%, the validation confusion matrix's prediction accuracy is almost 40%, and the test confusion matrix shows an accuracy of 39.8%. After training, validation, and testing, we collected 100,001 individual samples for overall performance evaluation.



Figure 8: All confusion matrix of diagnosis converter, DC-DC results by scaled conjugate gradient back propagation Neural Network.
Source: Authors (2026).

The diagonal cells of the confusion matrix represent the number of correctly classified instances for each class, while the off-diagonal cells indicate misclassified cases. The blue cell at the bottom right displays the overall accuracy (in green) and the total misclassification rate (in red). These results demonstrate strong recognition performance. Specifically, within the red circle, the neural network successfully identified 39.9% of fault events. In each confusion matrix: The top-left green cell shows the absolute number and percentage of true positive (TP) cases, representing correctly identified fault instances. The central green cell indicates the true negative (TN) cases, correctly identifying the absence of faults. The top red cell reflects false positive (FP) cases — instances incorrectly classified as faults (Type I error). The middle left red cell represents false negative (FN) cases — faults that were missed or classified as normal (Type II error).

This detailed breakdown provides insight into the network's classification strengths and areas for improvement. The top right gray square with green letters describes the positive predictive value [TP/(TP+FP)] ×100, and with the red letters the [FP/(TP+FP) ×100. The middle right gray square with green letters describes the negative predictive value [TN/(TN+FN)] ×100, and with the red letters the [FN/(TN+FN)] ×100. The bottom left gray square with green letters describes the sensitivity [TP/(TP+FN)] × 100, and the red letters describe the [FN/(TP+FN)] × 100. The bottom middle gray square with green letters describes the specificity [TN/(TN+FP)] × 100, and the red letters indicate the [FN/(TN+FP)] × 100. Finally, the bottom right blue square with green letters indicates the correctly classified incidents, [(TP+TN)/total incidents used] × 100, and the red letters indicate the incorrectly classified incidents, [(FP+FN)/total incidents used] × 100. The curve below gives us the overall performance of the network:

The y-axis represents the mean squared error, where a lower error indicates better performance. The x-axis demonstrates the epochs, which represent a step in the training of the NN. The best results were achieved at epoch 1000 in our NN.

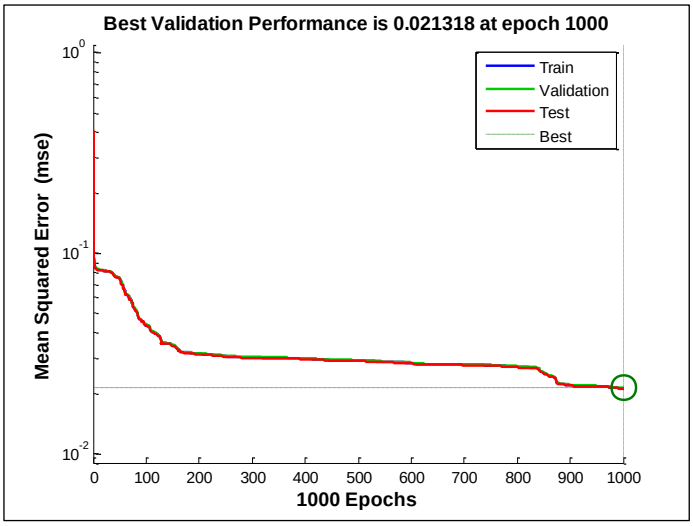


Figure 9: Performance of Neural Network.
Source: Authors (2026).

Figure 9 clearly shows that around 1000 epochs, the training, testing, and validation errors stabilize. After epoch 1000, the training error reaches a very low value of 0.0301, indicating a highly successful training process with a close match between actual and predicted outputs. The best validation performance is achieved at epoch 1000, with a mean squared error (MSE) of 0.0213. Initially, the MSE is relatively high, but it decreases steadily during training, reaching its minimum at epoch 1000. This demonstrates that the network's performance improves progressively with training, as further supported by Figure 3.5, where training and testing results closely align.

III.3 FAULTY CONDITION (ONLINE):

❖ The following figure represents the variation of the output voltage caused by faults.

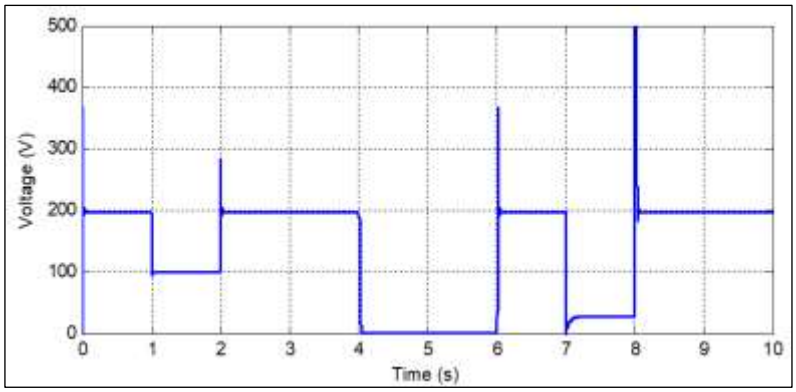


Figure 10: Fault output voltage.
Source: Authors (2026).

❖ The following figure represents the variation of the input current caused by the faults:

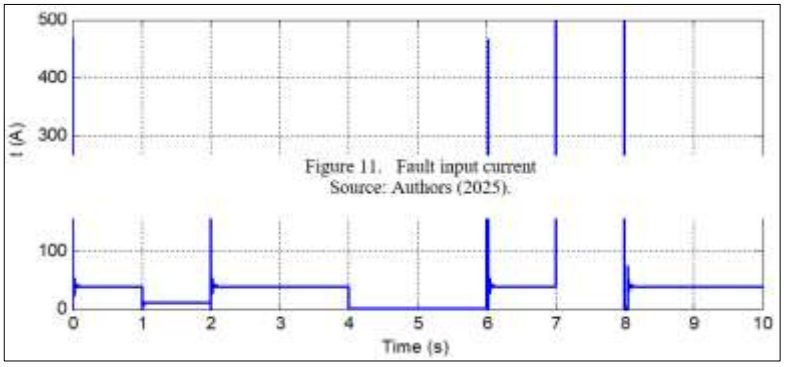


Figure 11. Fault input current
Source: Authors (2025).

Figure 11: Fault input current.
Source: Authors, (2026).

In the preceding figures, we clearly see the variation in output voltage and input current due to the faults generated in the converter components. These variations will be used later for the detection.

❖ The following figure represents the ANN fault indicator (Observing and indicating faults)

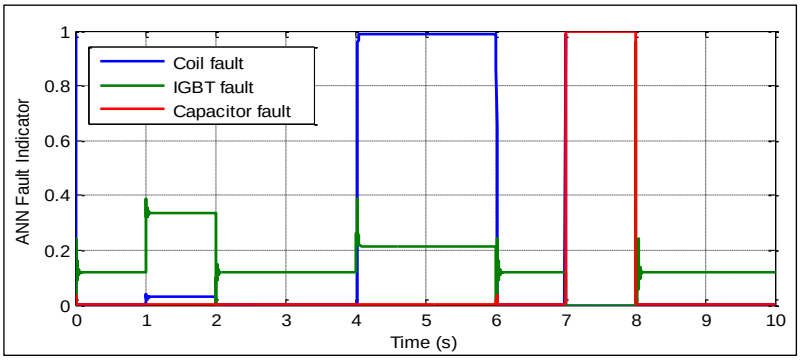


Figure 12: Online ANN fault indicator.
Source: Authors (2026).

The figure illustrates the performance of our fault smart indicator following the detection of defects. The diagnosis of serous neurons can easily locate the default by percentage signals. Product failure is most likely the one with the highest failure percentage (%).

IV. CONCLUSIONS

The purpose of this study is to apply an artificial neural network (ANN) technique for fault diagnosis in power converters. We aimed to address the challenge of online diagnosis by focusing on the identification and localization of component faults within DC-DC converters. Our approach involves implementing neural network structures tailored for this task. We reviewed basic topologies of DC-DC converters, including buck, boost, and buck-boost configurations. For each converter type, we analyzed the behavior of voltage and current through power switches, voltage gains, voltage and current ripples, and typical applications. We also surveyed existing fault diagnosis techniques for DC-DC converters.

An introduction to artificial neural networks was provided, covering fundamental concepts such as architecture, operational principles, learning mechanisms, and various types of neural networks and learning methods. We concluded with a discussion on their applications and reliability in challenging conditions.We detailed the key steps of our diagnosis system, utilizing a pattern recognition approach and selecting parameters validated numerically. Based on the obtained results, this diagnostic method has proven effective, especially in online operations.Following training, the ANN system can accurately detect and localize various faults in real time, supporting proactive maintenance and improving the reliability of the converter components.]

V. AUTHOR'S CONTRIBUTION

Conceptualization: Naas Djeddaoui.
Methodology: Naas Djeddaoui.
Investigation: Naas Djeddaoui.
Discussion of results: Naas Djeddaoui.
Writing – Original Draft: Naas Djeddaoui.
Writing – Review and Editing: Naas Djeddaoui.
Resources: Naas Djeddaoui.
Supervision: Naas Djeddaoui.
Approval of the final text: Naas Djeddaoui.

VI. REFERENCES

[1] A. Sumper and A. Baggini, Electrical Energy Efficiency: Technologies and Applications. Chichester, UK: John Wiley & Sons, 2012.

[2] V. Sharma and S. S. Chandel, "Performance and degradation analysis for long term reliability of solar photovoltaic systems: A review," Renewable and Sustainable Energy Reviews, vol. 27, pp. 753-767, 2013.

[3] M. M. R. Shamim and R. A. Ruddro, "Integration of PLC and smart diagnostics in predictive maintenance of CT tube manufacturing systems," International Journal of Scientific Interdisciplinary Research, vol. 1, no. 01, pp. 62-96, 2022.

[4] C. Gao et al., "Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease," Scientific Reports, vol. 8, no. 1, p. 7129, 2018.

[5] V. Venkatasubramanian et al., "A review of process fault detection and diagnosis: Part I: Quantitative model-based methods," Computers & Chemical Engineering, vol. 27, no. 3, pp. 293-311, 2003.

[6] R.-E. Precup, R.-C. Roman, and A. Safaei, Data-Driven Model-Free Controllers. Boca Raton, FL, USA: CRC Press, 2021.

[7] Z. Hu et al., "Deep neural networks with massive learned knowledge," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, USA, 2016, pp. 1670-1679.

[8] E. Jamshidpour et al., "Single-switch DC–DC converter with fault-tolerant capability under open-and short-circuit switch failures," IEEE Transactions on Power Electronics, vol. 30, no. 5, pp. 2703-2712, May 2014.

[9] D. R. Espinoza Trejo et al., "Open-and short-circuit fault identification for a boost dc/dc converter in PV MPPT systems," Energies, vol. 11, no. 3, p. 616, 2018.

[10] MathWorks, "Pattern Recognition," 2024. [Online]. Available: https://www.mathworks.com/discovery/pattern-recognition.html. Accessed: May 17, 2024..

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 323-331. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2927

RESEARCH ARTICLE OPEN ACCESS

ENHANCING SWELLING PRESSURE PREDICTION THROUGH INTEGRATED MACHINE LEARNING METHODS

Laid Lekouara[1], Fatima Zohra Tebbi[2], Ouassila Bahloul[3] and Rachid Rabehi[*4]

[1]Faculty of Science and Technology, Department of Civil Engineering, University Abbes Laghrour, Khenchela, Algeria.
[2]LRNAT Laboratory, Earth Sciences and Universe Institute. University of Mostepha Ben Boulaid, Batna 2. Algeria.
[3]Civil Engineering Laboratory – Risks and Structures in Interactions,Department of Civil Engineering,Faculty of Technology, University of Batna 2 – Mostefa Ben Boulaïd Batna 05000, Algeria.
[4]Faculty of Civil Engineering, University of Science and Technology Houari Boumediene (USTHB), Bab Ezzouar, 16111, Algeria.

[1]https://orcid.org/0009-0006-6700-1245, [2]https://orcid.org/0000-0003-3321-5609
[3]https://orcid.org/0000-0001-8554-7563, [4]https://orcid.org/0000-0002-9398-7291

Email: lekouara_laid@univ-khenchela.dz, f.tebbi@univ-batna2.dz, o.bahloul@univ-batna2.dz, rachid.rabehi@usthb.edu.dz

ARTICLE INFO

Article History
Received: November 11, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Swelling pressure,
Machine learning,
Feature selection,
Random forest,
Neural network.

ABSTRACT

Swelling soils, characterized by their high capacity to change volume due to variations in water content, pose significant challenges in civil engineering, including structural deformations, cracks in infrastructure, and foundation instabilities. These issues lead to high maintenance costs and safety risks for structures. To address these challenges, this study proposes an integrated approach combining machine learning and feature selection techniques to predict the swelling pressure (SP, kPa) of soils. Using a dataset comprising seven explanatory variables (liquid limit (LL, %), plasticity index (PI, %), shrinkage limit (WS, %), particle size fractions (Ff2mm, Ff80µm, Ff2µm), and methylene blue index (VBS, g/100 g)), four feature selection methods were evaluated: Least Absolute Shrinkage and Selection Operator (Lasso), Random Forest, Gradient Boosting, and the F-statistic test. These methods identified the five most influential variables for training an artificial neural network (ANN). The results show that Random Forest achieved the best predictive performance (MSE = 785.31, R^2 = 0.9706, MAE = 21.39), followed by the F-test (MSE = 1784.99, R^2 = 0.9332, MAE = 30.66), Gradient Boosting (MSE = 2497.38, R^2 = 0.9066, MAE = 40.20), and Lasso CV (MSE = 2605.78, R^2 = 0.9035, MAE = 40.29). Analysis of the variable distributions revealed complex nonlinear relationships between geotechnical parameters and swelling pressure. This multi-method approach demonstrates the effectiveness of machine learning techniques for accurate and cost-effective prediction of geotechnical properties, offering a promising alternative to traditional, time-consuming, and costly oedometer tests.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Expansive soils, characterized by their significant volumetric changes in response to moisture content variations, represent one of the most challenging geotechnical phenomena in civil engineering practice [1],[2]. These soils, predominantly composed of clay minerals with high swelling potential, pose substantial threats to infrastructure stability, causing billions of dollars in damage annually through structural deformations, foundation failures, and pavement distress [3],[4]. The swelling pressure generated by these soils upon moisture increase can reach magnitudes sufficient to cause significant uplift forces on foundations and lateral pressures on retaining structures, necessitating accurate prediction methods for safe and economical design [5],[6]. However, traditional approaches to swelling pressure determination rely primarily on laboratory oedometer tests which, while providing reliable results, are time-consuming, expensive, and require specialized equipment and expertise [7],[8]. These limitations have driven the geotechnical engineering community to seek alternative prediction methods based on index properties such as liquid limit (LL), plasticity index (PI), shrinkage limit (WS), particle

size fractions (Ff2mm, Ff80µm, Ff2µm), and methylene blue index (VBS), but their applicability often remains limited to specific soil types [9-11]. In this context, the advent of artificial intelligence and machine learning has revolutionized numerous engineering disciplines, with geotechnical engineering particularly benefiting from these computational advances [12-14]. Artificial neural networks (ANNs), in particular, have emerged as powerful tools for modeling complex, nonlinear relationships inherent in geotechnical problems, due to their ability to capture intricate patterns in data without requiring explicit mathematical formulations [15-19]. Furthermore, recent studies have demonstrated the superior performance of machine learning methods in swelling pressure prediction. According to [20] successfully applied genetic expression programming and artificial neural networks to predict expansive soil properties. In [21] compared multiple machine learning models, demonstrating that neural networks consistently outperformed traditional statistical methods with R^2 values exceeding 0.92.

Similarly, [22] developed constrained XGBoost models achieving remarkable accuracy with R^2 values of 0.9832, while [23] demonstrated the effectiveness of dimensionality reduction combining linear (PCA) and nonlinear (ISOMAP) techniques with neural networks for swelling pressure prediction in Northeast Algeria. Moreover, feature selection plays a crucial role in machine learning applications, particularly in geotechnical engineering where numerous correlated parameters influence soil behavior [24],[25]. Various feature selection techniques have been employed, with particular focus on methods such as Least Absolute Shrinkage and Selection Operator (Lasso), Random Forest importance ranking, Gradient Boosting feature scoring, and statistical F-test analysis. In particular, advanced algorithms such as Boruta have shown promise, as demonstrated by [26] in their model for the Mila region. Lasso regression (LassoCV), known for its automatic feature selection capability through L1 regularization, has been successfully applied in geotechnical applications [27], while Random Forest and Gradient Boosting methods have demonstrated superior performance in capturing nonlinear relationships and feature interactions [28],[29].

Nevertheless, statistical approaches, including F-test based feature selection, continue to provide valuable insights into variable significance, although their linear assumptions may limit effectiveness with highly nonlinear geotechnical data [30],[31]. However, when combined with nonlinear modeling approaches such as ANNs, these methods can contribute to comprehensive predictive frameworks that leverage both statistical rigor and machine learning flexibility [32],[33]. Currently, machine learning applications in geotechnical engineering have expanded rapidly, with artificial neural networks representing approximately 52% of all AI applications in the field (Application of artificial intelligence in geotechnical engineering [34]). The versatility of neural networks in handling complex multidimensional problems has made them particularly attractive for soil behavior modeling [35-37], while support vector machines, random forest, and gradient boosting methods collectively represent nearly 40% of applications (State-of-the-art review on AI-enhanced computational mechanics, [38]). Nevertheless, current challenges include data quality and availability, model interpretability, and generalization to different soil types and environmental conditions [39-41].

Consequently, the integration of multiple machine learning approaches for feature selection and prediction represents a significant advancement in geotechnical engineering practice. Recent research has shown that ensemble methods and hybrid approaches often outperform individual techniques [42-44]. This integrated approach not only improves prediction accuracy through optimal feature subset identification but also provides insights into the relative importance of different geotechnical parameters, contributing to enhanced understanding of expansive soil behavior and the complex relationships between soil index properties and swelling pressure mechanisms. Furthermore, the practical implementation of machine learning-based swelling pressure prediction methods requires careful consideration of model validation using performance metrics such as Mean Squared Error (MSE), coefficient of determination (R^2), and Mean Absolute Error (MAE), along with uncertainty quantification and integration with existing geotechnical design practices [45],[46]. Moreover, recent advances in model interpretability and explainable AI have made machine learning approaches more acceptable to practicing engineers who require understanding of the underlying reasoning behind predictions [47-49]. Ultimately, the significance of this research extends to practical applications in geotechnical engineering design and construction.

Accurate swelling pressure prediction enables engineers to make informed decisions regarding foundation design, ground improvement strategies, and risk mitigation measures, contributing to the development of safer and more economical infrastructure in areas affected by expansive soils. Therefore, the main objective of this research is to develop and systematically compare four distinct feature selection methodologies - Least Absolute Shrinkage and Selection Operator (LassoCV), Random Forest, Gradient Boosting, and statistical F-test - coupled with artificial neural networks to identify the optimal combination for swelling pressure prediction, by rigorously evaluating their effectiveness on a comprehensive dataset of seven explanatory geotechnical variables including Atterberg limits (liquid limit LL and plasticity index PI), shrinkage limit (WS), particle size fractions (Ff2mm, Ff80µm, Ff2µm), and methylene blue index (VBS), using robust performance metrics (MSE, R^2, MAE) and providing evidence-based practical guidelines for method selection in geotechnical engineering applications.

II. MATERIALS AND METHODS

II.1 DATASET

The dataset comprises 160 observations with eight parameters: LL, PI, WS, Ff2mm, Ff80, Ff2, VBS, and SP. Descriptive statistics are provided in Table 1. The target variable, SP, has a mean of 616.82 kPa, a standard deviation of 136.24 kPa, and ranges from 389 to 936 kPa, reflecting significant variability in swelling pressure.

Table 1: Descriptive statistics for all parameters.

Parameter	Mean	Std Dev	Median	Min	Max	Skewness
LL	80.78	7.86	81.71	65.78	95.30	-0.33
PI	51.83	4.97	51.93	41.66	66.97	0.09
WS	8.88	1.79	8.60	5.80	12.60	0.24
Ff2mm	97.08	1.53	97.33	92.54	99.37	-0.85
Ff80	90.06	5.70	91.35	70.80	97.29	-1.30
Ff2	49.63	7.48	50.63	30.10	62.00	-0.70

VBS	12.53	2.09	12.58	8.00	16.00	-0.13
SP	612.82	136.24	582.00	389.00	936.00	0.74

Source: Authors, (2026).

III. METHODOLOGY

The methodology employed for predicting soil pressure (SP) followed a structured pipeline designed to optimize feature selection and neural network (NN) performance, as depicted in the provided flowchart (Figure 1). The process began with the loading of experimental data, containing 160 samples with seven initial features (LL, PI, WS, Ff2mm, Ff80µm, Ff2µm, VBS) and the target variable SP. All analyses were conducted using Python 3 in Google Colab, providing a cloud-based computational environment with access to necessary machine learning libraries and GPU acceleration capabilities. Data preparation involved handling missing values by imputing feature means, splitting the dataset into training (80%) and testing (20%) sets, and standardizing both features and the target variable using StandardScaler. Four feature selection methods—LassoCV, Random Forest, Gradient Boosting, and Statistical F-test—were applied to identify the top five most relevant features based on their respective criteria: absolute coefficients for Lasso CV, feature importance for Random Forest and Gradient Boosting, and F-scores for the F-test. Each selected feature subset was used to train a dedicated NN model with a multi-layer architecture (128-64-32-16-1 neurons with dropout layers) using the Adam optimizer, incorporating early stopping with a patience of 20 epochs to prevent overfitting. The models were evaluated using Mean Squared Error (MSE), R^2 score, and Mean Absolute Error (MAE), with training conducted over a maximum of 200 epochs and a batch size of 16. Finally, a comprehensive comparison of the models' performance was conducted, including bar plots for R^2, MSE, and feature counts.

III.1 MACHINE LEARNING APPROACHES

Four feature selection methods were applied to identify the five most relevant variables. The algorithms for each method are described below:

III.1.1 Algorithm Lasso Selection

1) Let M0 denote the null model, which contains no predictors and predicts the mean of the target (e.g SP).

2) For k = 0, ……, p – 1.
a. Consider all models that augment the predictors in Mk with a regularization parameter α.
b. Fit a Lasso regression model to minimize the cost function:

$$Cost = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha \sum_{j-1}^{p}|\beta_j|. \quad (1)$$

Where y_i is the observed target, $\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_J \, x_{ij}$ is the prediction, n is the number of samples, p is the number of predictors, α is the regularization parameter, and $\sum_{j-1}^{p}|\beta_j|$ is the L1 penalty.

c. Choose the best model among these by selecting the α that optimizes a cross validated error metric (e.g, MSE).
d. Update $M_{(k+1)}$ with predictors whose coefficients $\beta_j \neq 0$, indicating their selection.

3) Select a single best model from among M_0, ……, M_p using cross validated prediction error (e.g., C_P, AIC, BIC, or adjusted R^2).

4) Output the final model with the selected predictors and their corresponding coefficients.

III.1.2 Algorithm Lasso Selection

1) Let M_0 denote the null model, which contains no predictors and predicts the mean of the target (e.g SP).

2) For k = 0, ……, p – 1.
a. Consider all p – k models that augment the predictors in M_k with a subset of the remaining predictors.
b. Fit a Random Forest model to the training data, using a specified number of trees (e.g., 200) and a Random subset of predictors at each split.
c. Compute the feature importance scores based on the decrease in node impurity across all trees.
d. Choose the best model among these by selecting the top N predictors with the highest importance scores, and update M_{k+1} with these predictors.

3) Select a single best model from among M_0, ……, M_p using cross validated prediction error (e.g., C_P, AIC, BIC, or adjusted R^2).

4) Output the final model with the selected predictors and their importance scores.

III.1.3 Algorithm Gradient Boosting Selection

1) Let M_0 denote the null model, which contains no predictors and predicts the mean of the target (e.g SP).

2) For k = 0, ……, p – 1.
a. Consider all p – k models that augment the predictors in M_k with a subset of the remaining predictors.
b. Fit a Gradient Boosting model to the training data, iteratively adding weak learners (e.g., decision trees) to minimize the loss function by fitting to the residuals.

c. Compute the feature importance scores based on the contribution of each predictor to reducing the loss across all iterations.
d. Choose the best model among these by selecting the top N predictors with the highest importance scores, and update M_{k+1} with these predictors.

3) Select a single best model from among M_0, ……, M_p using cross validated prediction error (e.g., CP, AIC, BIC, or adjusted R^2).

4) Output the final model with the selected predictors and their importance scores.

III.1.4 Algorithm f-Test Selection

1) Let M_0 denote the null model, which contains no predictors and predicts the mean of the target (e.g SP).

2) For k = 0, ……, p – 1.
a. Consider all p – k models that augment the predictors in Mk with one additional predictors.
b. Fit a linear regression model to the training data, for each model and compute the F-statistic for each additional predictor, measuring its significance in explaining the target (e.g SP).
c. Compute the F-scores and associated p-values for each predictor based onthe F-statistic.
d. Choose the best model among these by selecting the top N predictors with the highest importance scores, and update M_{k+1} with these predictors.

3) Select a single best model from among M_0, ……, M_p using cross validated prediction error (e.g., C_P, AIC, BIC, or adjusted R^2).

4) Output the final model with the selected predictors and their importance scores.

III.1.5 Algorithm Neural Network Predictor

1) Let M_0 denote the initial model, which is an untrained neural network with a predefined architecture and no learned weights.

2) For each training epoch t = 1…. T:
a. Consider the current model M_{t-1} and prepare the training data (e.g., $X_{\text{train-selected}}$, $Y_{\text{train-scaled}}$).
b. Forward propagate the input data through the network layers to compute predictions and calculate the loss (e.g MSE) against the target (e.g SP).
c. Backpropagate the error using gradient descent (e.g., Adam optimizer) to update the network weights.
d. Evaluate the validation loss on a held-out subset, and if it does not improve for a specified number of epochs, stop training and restore the best weights.

3) Select the final model MT based on the minimum validation loss achieved during training.

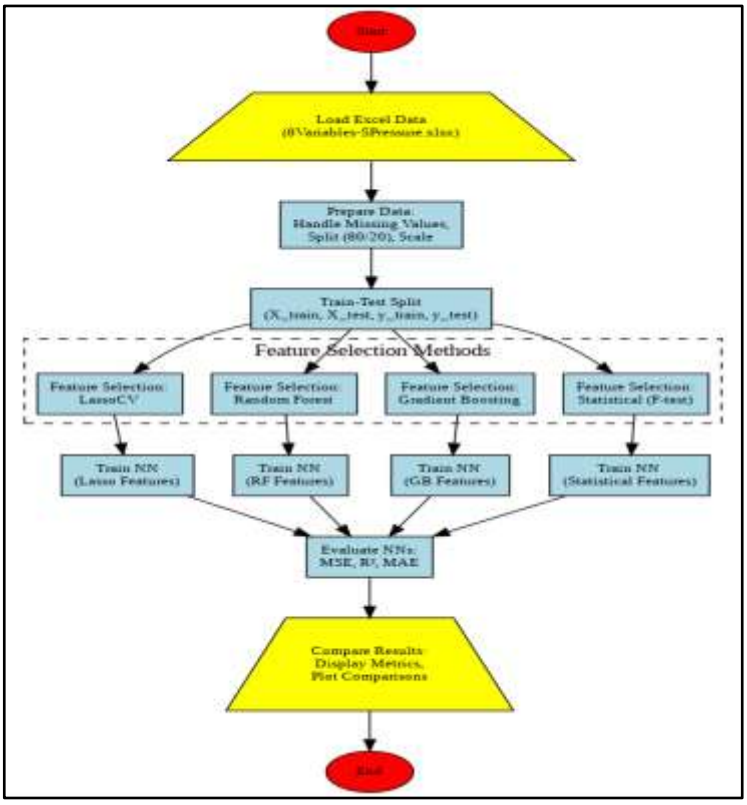


Figure 1: Flowchart of the modeling process.
Source: Authors, (2026).

III.2 PERFORMANCE METRICS

The performance of each model was evaluated using three:

III.2.1 Erreur Quadratique Moyenne (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

y_i : Actual (or observed) value for the i-th observation. This is target value the model aims to predict.
$\hat{y}_i$: Predicted value by the model for the i-th observation. This is the model's output.

III.2.2 Coefficient de Détermination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

y_i : Actual value for the i-th observation
$\hat{y}_i$: Predicted value for the i-th observation
$\bar{y}$: Mean of the acyual values.

III.2.3 Erreur Absolue Moyenne (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \quad (4)$$

y_i : Actual value for the i-th observation
$\hat{y}_i$: Predicted value for the i-th observation
$\bar{y}$: Total number of observations, as in MSE.

IV. RESULTS

The comparative evaluation of the four methods reveals a significant performance hierarchy for swelling pressure prediction. All methods select exactly 5 parameters out of 7, but the high-performing methods (Random Forest, Gradient Boosting) prioritize fine fractions (Ff2, Ff80) that are essential to clay behavior, unlike Lasso which selects the less relevant coarse fraction (Ff2mm), explaining its inferior performance (Figure 2). Random Forest demonstrates exceptional superiority with a coefficient of determination of 0.9706 (MSE = 785.31 kPa², MAE = 21.39 kPa), explaining 97.06% of the total variance in swelling pressure. This performance stems from its unique algorithmic capacity to capture highly complex nonlinear relationships between the selected critical geotechnical variables (Ff2, LL, Ff80, PI, VBS). The superiority rests on a sophisticated impurity reduction mechanism based on the Gini criterion that enables remarkable precision in identifying the most discriminating variables, even in the presence of complex multivariate interactions and partial collinearities between geotechnical parameters, through its ensemble approach averaging predictions from hundreds of decision trees trained on different bootstrap subsamples (Figure 3 and 7).

The statistical F-test achieves solid performance in second position with robust metrics (R^2 = 0.9332, MSE = 1784.99 kPa², MAE = 30.66 kPa), explaining 93.32% of the variance. This performance testifies to the persistent effectiveness of classical statistical approaches in modern geotechnical analysis. However, the specific inclusion of WS (shrinkage limit) instead of Ff80 (particle size fraction < 80µm) slightly reduces its predictive accuracy compared to Random Forest, revealing a fundamentally different selection philosophy based on identifying direct linear relationships between individual parameters and swelling pressure (Figure 4 and 7). Gradient Boosting presents a methodological paradox with moderate performance (R^2 = 0.9066, MSE = 2497.38 kPa², MAE = 40.20 kPa) despite feature selection nearly identical to Random Forest (LL, Ff2, Ff80, PI, VBS). This illustrates an algorithmic paradox where its theoretically powerful sequential adaptive learning capability proves more sensitive to overfitting and stochastic variations than Random Forest's more stable ensemble approach.

The increased sensitivity is explained by the sequential nature of learning that can amplify modeling errors and higher variance due to dependency between successive predictors (Figure 5 and 7). Lasso CV displays the weakest performance with R^2 = 0.9035, MSE = 2605.78 kPa², and MAE = 40.29 kPa, primarily due to the suboptimal inclusion of Ff2mm (fraction < 2mm) instead of Ff80 (fraction < 80µm). This deficient selection reveals the fundamental limitations of L1 regularization which, through its rigid binary selection mechanism and strictly linear mathematical assumptions, cannot effectively capture the complex physicochemical interactions characteristic of clay swelling behavior. The crystalline (interlayer spacing), osmotic (ionic swelling pressure), and capillary (matric suction) processes interact in a highly synergistic, multiplicative, and non-additive manner. These complex mechanisms require sophisticated machine learning approaches capable of transcending the limitations of traditional linear models to exploit the informational richness of nonlinear patterns hidden in geotechnical data (Figure 6 and 7).

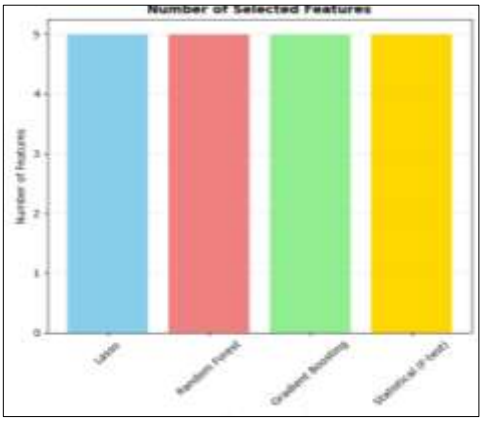


Figure 2: Selected features by each feature selection method.
Source: Authors, (2026).

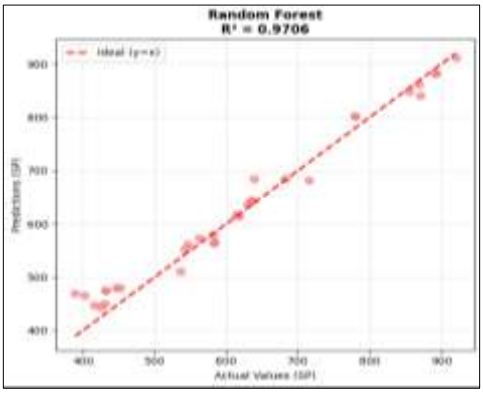


Figure 3: Random Forest performance for swelling pressure prediction.
Source: Authors, (2026).

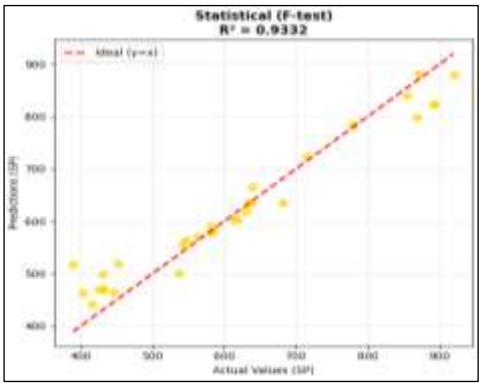


Figure 4: Staitistical (F-test) performance for swelling pressure prediction.
Source: Authors, (2026).

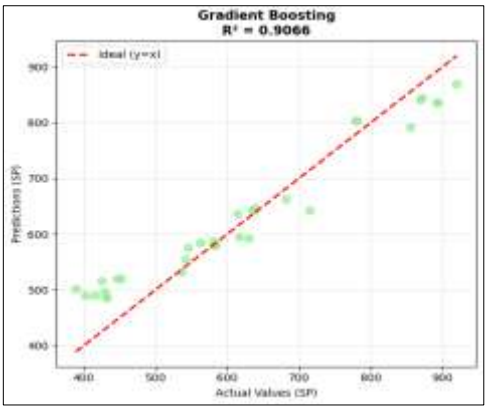


Figure 5: Gradient Boosting performance for swelling pressure prediction.
Source: Authors, (2026).

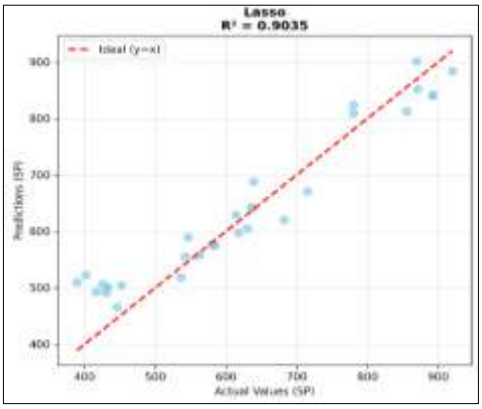


Figure 6: Lasso performance for swelling pressure prediction.
Source: Authors, (2026).

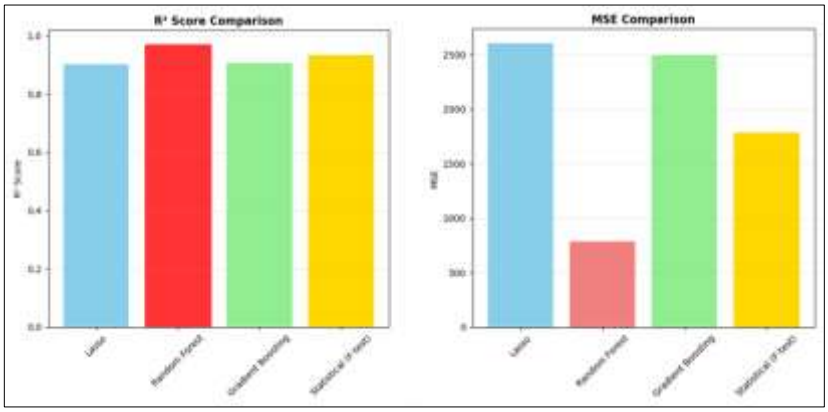


Figure 7: Comparative performance evaluation of feature selection methods using R² and MSE metrics.
Source: Authors, (2026).

IV CONCLUSIONS

The comparative analysis reveals significant performance differences between feature selection methods coupled with neural networks, Random Forest demonstrating clear superiority with an R^2 of 0.9706 using features Ff2, LL, Ff80, PI and VBS, highlighting the crucial importance of selection technique choice for optimizing predictive accuracy. This exceptional performance confirms the remarkable capacity of ensemble methods to capture complex nonlinear relationships inherent in clay swelling behavior, particularly through precise identification of critical fine particle size fractions. The statistical F-test achieves robust results (R^2 = 0.9332), demonstrating that classical statistical approaches retain considerable relevance, although the observed paradox with Gradient Boosting (R^2 = 0.9066) illustrates that optimal feature selection does not automatically guarantee superior performance, revealing the fundamental importance of algorithmic stability. The results reveal that all methods systematically select five parameters among the seven available variables, but high-performing techniques favor fine fractions essential to the complex physicochemical mechanisms governing clay swelling.

Future work should prioritize refinement of selection methods and architectural optimization of models. The Random Forest-neural network integration is recommended for preliminary swelling pressure evaluation, offering optimal balance between predictive accuracy, computational efficiency and result interpretability, facilitating its adoption by geotechnical engineering practitioners as a complement to traditional tests for critical applications. This multi-method approach represents a paradigmatic shift toward data-driven geotechnics, establishing foundations for future research in expansive soil characterization and encouraging broader adoption of artificial intelligence in geotechnics, ultimately contributing to the development of safer and more economical infrastructure by enabling engineers to make informed decisions based on robust scientific evidence, thereby reducing risks and optimizing costs in infrastructure projects affected by expansive soils.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Laid Lekouara, Fatima Zohra Tebbi, Ouassila Bahloul and Rachid Rabehi.
Methodology: Laid Lekouara, Fatima Zohra Tebbi, Ouassila Bahloul and Rachid Rabehi.
Investigation: Laid Lekouara, Fatima Zohra Tebbi, Ouassila Bahloul and Rachid Rabehi.
Discussion of results: Laid Lekouara, Fatima Zohra Tebbi, Ouassila Bahloul and Rachid Rabehi.
Writing – Original Draft: Laid Lekouara, Fatima Zohra Tebbi, Ouassila Bahloul and Rachid Rabehi.
Writing – Review and Editing: Laid Lekouara, Fatima Zohra Tebbi, Ouassila Bahloul and Rachid Rabehi.
Resources: Laid Lekouara, Fatima Zohra Tebbi, Ouassila Bahloul and Rachid Rabehi.
Supervision: Laid Lekouara, Fatima Zohra Tebbi, Ouassila Bahloul and Rachid Rabehi.
Approval of the final text: Laid Lekouara, Fatima Zohra Tebbi, Ouassila Bahloul and Rachid Rabehi.

VI. REFERENCES

[1] V. Cantillo, V. Mercado, C. Pájaro, Empirical correlations for the swelling pressure of expansive clays in the city of Barranquilla, Colombia, Earth Sciences Research Journal 21 (2017) 45–49. https://doi.org/10.15446/esrj.v21n1.60226.

[2] C.I. Christopher, D.N. Chimobi, Emerging trends in expansive soil stabilisation: A review, Journal of Rock Mechanics and Geotechnical Engineering 11 (2019) 423–440. https://doi.org/10.1016/j.jrmge.2018.08.013.

[3] A. Khennouf, M. Baheddi, Heave analysis of shallow foundations founded in swelling clayey soil at N'Gaous city in Algeria, Studia Geotechnica et Mechanica 42 (2020) 210–221. https://doi.org/10.2478/sgem-2020-0009.

[4] J. Du, S. Zhang, H. Li, X. Zhao, L. Wang, Characterization of controlled low-strength materials from waste expansive soils, Construction and Building Materials 411 (2024) 134690. https://doi.org/10.1016/j.conbuildmat.2023.134690.

[5] B. Sharmila, S. Bhuvaneshwari, G. Landlin, Application of lignosulphonate: A sustainable approach towards strength improvement and swell management of expansive soils, Bulletin of Engineering Geology and the Environment 80 (2021) 6395–6413. https://doi.org/10.1007/s10064-021-02332-8.

[6] M.A. Waheed, O.S.B. Al-Amoudi, M.A. Al-Osta, Molecular-level behavior induction in the constitutive modeling of swelling clayey soils: A review, Transportation Geotechnics 39 (2023) 100947. https://doi.org/10.1016/j.trgeo.2023.100947.

[7] Z.A. Erguler, R. Ulusay, A simple test and predictive models for assessing swell potential of Ankara (Turkey) clays, Engineering Geology 67 (2003) 331–352. https://doi.org/10.1016/S0013-7952(02)00205-3.

[8] V.H.R. Barbosa, M.E.S. Marques, A.C.R. Guimarães, C.D. Castro, Characterization of an expansive soil in southwest Brazilian amazon—behavior of an expansive subgrade in a flexible pavement, In Advances in Transportation Geotechnics IV, Springer (2022) 323–335. https://doi.org/10.1007/978-3-030-77234-5_24.

[9] A.A. Basma, Prediction of expansion degree for natural compacted clays, Geotechnical Testing Journal 16 (1993) 542–549. https://doi.org/10.1520/GTJ10291J.

[10] I. Ashayeri, S. Yasrebi, Free-swell and swelling pressure of unsaturated compacted clays; experiments and neural networks modeling, Geotechnical and Geological Engineering 27 (2009) 137–153. https://doi.org/10.1007/s10706-008-9210-7.

[11] K. Meshram, N. Singh, P. Jain, Estimation of swelling characteristics of expansive soils with influence of clay mineralogy, Applied Clay Science 198 (2021) 105822. https://doi.org/10.1016/j.clay.2020.105822.

[12] M.A. Shahin, State-of-the-art review of some artificial intelligence applications in pile foundations, Geoscience Frontiers 7 (2016) 33–44. https://doi.org/10.1016/j.gsf.2014.10.006.

[13] S. Jong, D. Ong, E. Oh, State-of-the-art review of geotechnical-driven artificial intelligence techniques in underground soil–structure interaction, Tunnelling and Underground Space Technology 113 (2021) 103946. https://doi.org/10.1016/j.tust.2021.103946.

[14] W. Zhang, X. Gu, L. Hong, L. Han, L. Wang, Comprehensive review of machine learning in geotechnical reliability analysis: Algorithms, applications and further challenges, Applied Soft Computing 136 (2023) 110066. https://doi.org/10.1016/j.asoc.2023.110066.

[15] A.T.C. Goh, Nonlinear modelling in geotechnical engineering using neural networks, Transactions of the Institution of Engineers, Australia. Civil Engineering 36 (1994) 293–297.

[16] Y.M. Najjar, H.E. Ali, CPT-based liquefaction potential assessment: A neuronet approach, Geotechnical Special Publication 75 (1998) 542–553.

[17] M. Lefik, Artificial neural network applications for geotechnical engineering, Applied Sciences 13 (2023) 2988. https://doi.org/10.3390/app13052988.

[18] J.J. Jeremiah, S.J. Abbey, C.A. Booth, A. Kashyap, Results of application of artificial neural networks in predicting geo-mechanical properties of stabilised clays—a review, Geotechnics 1 (2021) 147–171. https://doi.org/10.3390/geotechnics1010008.

[19] Q. Guan, Z. Yang, N. Guo, Z. Hu, Finite element geotechnical analysis incorporating deep learning-based soil model, Computers and Geotechnics 154 (2023) 105120. https://doi.org/10.1016/j.compgeo.2022.105120.

[20] F.E. Jalal, Y. Xu, M. Iqbal, M.F. Javed, B. Jamhiri, Predictive modeling of swell-strength of expansive soils using artificial intelligence approaches: ANN, ANFIS and GEP, Journal of Environmental Management 289 (2021) 112420. https://doi.org/10.1016/j.jenvman.2021.112420.

[21] S. Gahlot, R. Mangal, A. Arya, B.P. Sethy, K.C. Sethi, Prediction of swelling pressure of expansive soil using machine learning methods, Asian Journal of Civil Engineering 25 (2024) 3279–3292. https://doi.org/10.1007/s42107-024-01205-7.

[22] M. Shehab, R. Taherdangkoo, C. Butscher, Towards reliable barrier systems: A constrained XGBoost model coupled with gray wolf optimization for maximum swelling pressure of bentonite, Computers and Geotechnics 168 (2024) 106132. https://doi.org/10.1016/j.compgeo.2024.106132.

[23] O. Bahloul, F.Z. Tebbi, L. Lekouara, H. Bekhouche, Neural networks based linear (PCA) and nonlinear (ISOMAP) feature extraction for soil swelling pressure prediction (North East Algeria), Heliyon 9 (2023) e18673. https://doi.org/10.1016/j.heliyon.2023.e18673.

[24] I.G. Tejada, P. Antolin, Use of machine learning for unraveling hidden correlations between particle size distributions and the mechanical behavior of granular materials, Acta Geotechnica 17 (2022) 1443–1461. https://doi.org/10.1007/s11440-021-01327-x.

[25] N. Tiwari, N. Satyam, Coupling effect of pond ash and polypropylene fiber on strength and durability of expansive soil subgrades: An integrated experimental and machine learning approach, Journal of Rock Mechanics and Geotechnical Engineering 13 (2021) 1101–1112. https://doi.org/10.1016/j.jrmge.2021.03.010.

[26] L. Lekouara, O. Bahloul, F.Z. Tebbi, R. Rabehi, A swelling pressure prediction model based on boruta feature selection mila region, Studies in Engineering and Exact Sciences 5 (2024) e11613. https://doi.org/10.54021/seesv5n2-633.

[27] S.D. Mohammadzadeh, S.-F. Kazemi, A. Mosavi, E. Nasseralshariati, J.H.M. Tah, Prediction of compression index of fine-grained soils using a gene expression programming model, Infrastructures 4 (2019) 26. https://doi.org/10.3390/infrastructures4020026.

[28] M. Kumar, P. Samui, D.R. Kumar, P.G. Asteris, State-of-the-art XGBoost, RF and DNN based soft-computing models for PGPN piles, Geomechanics and Geoengineering 19 (2024) 320–335. https://doi.org/10.1080/17486025.2024.2307746.

[29] J. Khatti, K.S. Grover, Prediction of compaction parameters for fine-grained soil: Critical comparison of the deep learning and standalone models, Journal of Rock Mechanics and Geotechnical Engineering 15 (2023) 3010–3038. https://doi.org/10.1016/j.jrmge.2023.02.016.

[30] M. Kumor, Selected geotechnical problems of expansive clays in the area of Poland, Architecture Civil Engineering Environment 1 (2008) 75–92.

[31] R. Bag, A. Rabbani, Effect of temperature on swelling pressure and compressibility characteristics of soil, Applied Clay Science 136 (2017) 1–7. https://doi.org/10.1016/j.clay.2016.11.014.

[32] L. Yunlong, S. Vanapalli, Pile behavior modeling in unsaturated expansive soils, In Modeling in Geotechnical Engineering, Elsevier (2021) 393–427. https://doi.org/10.1016/B978-0-12-821205-9.00018-7.

[33] X. Chen, X. Shi, J. Zhou, X. Du, Q. Chen, X. Qiu, Effect of overflow tailings properties on cemented paste backfill, Journal of Environmental Management 235 (2019) 133–144. https://doi.org/10.1016/j.jenvman.2019.01.040.

[34] Application of artificial intelligence in geotechnical engineering, Geotechnical Engineering Research 40 (2022) 152–168.

[35] Y.M. Najjar, H.E. Ali, I.A. Basheer, On the use of neurons for simulating the stress-strain behavior of soils, Proceedings of the 7th International Conference on Neural Information Processing, Boulder, Colorado (1999).

[36] D. Penumadu, R. Zhao, Triaxial compression behavior of sand and gravel using artificial neural networks (ANN), Computers and Geotechnics 24 (1999) 207–230. https://doi.org/10.1016/S0266-352X(99)00002-4.

[37] G.W. Ellis, C. Yao, R. Zhao, D. Penumadu, Stress-strain modelling of sands using artificial neural networks, Journal of Geotechnical Engineering, ASCE 121 (1995) 429–435. https://doi.org/10.1061/(ASCE)0733-9410(1995)121:5(429).

[38] State-of-the-art review on AI-enhanced computational mechanics, Computational Mechanics and Engineering 45 (2024) 89–112.

[39] R. Taherdangkoo, A. Tatomir, T. Anighoro, M. Sauter, Modeling of hydraulic fracturing process in unconventional reservoirs: A review, Environmental Earth Sciences 82 (2023) 1–23. https://doi.org/10.1007/s12665-023-10803-1.

[40] P. Narmandakh, R. Taherdangkoo, M. Javadi, C. Butscher, Neural network model to determine maximum swelling pressure of clayey soils, Computers and Geotechnics 154 (2023) 105157. https://doi.org/10.1016/j.compgeo.2022.105157.

[41] M. Ozsagir, C. Erden, E. Bol, S. Sert, A. Özocak, Machine learning approaches for prediction of fine-grained soils liquefaction, Computers and Geotechnics 152 (2022) 105014. https://doi.org/10.1016/j.compgeo.2022.105014.

[42] M. Habib, B. Bashir, A. Alsalman, H. Bachir, Prediction and parametric assessment of soil one-dimensional vertical free swelling potential using ensemble machine learning models, Advanced Modeling and Simulation in Engineering Sciences 11 (2024) 26. https://doi.org/10.1186/s40323-024-00270-w.

[43] B. Alibrahim, A.H.B. Garoushi, E. Uygar, The role of calcium-based additives in bentonite stabilization: a comparative evaluation, Arabian Journal of Science and Engineering 49 (2024) 7751–7764. https://doi.org/10.1007/s13369-024-09281-2.

[44] A.A. Sabtan, Geotechnical properties of expansive clay shale in Tabuk, Saudi Arabia, Journal of Asian Earth Sciences 25 (2005) 747–757. https://doi.org/10.1016/j.jseaes.2004.07.003.

[45] X. Guo, H. Shi, W.A. Dick, L. Ping, X. Chen, Modeling the influence of lime on the unconfined compressive strength of reconstituted graded soil using advanced machine learning approaches for subgrade and liner applications, PLoS ONE 19 (2024) e0301075. https://doi.org/10.1371/journal.pone.0301075.

[46] S. Hanandeh, A. Ardah, M. Abu-Farsakh, Using artificial neural network and genetics algorithm to estimate the resilient modulus for stabilized subgrade and propose new empirical formula, Transportation Geotechnics 24 (2020) 100358. https://doi.org/10.1016/j.trgeo.2020.100358.

[47] M.R.H. Hossain, M.A. Kabir, Machine learning techniques for estimating soil moisture from smartphone captured images, Agriculture 13 (2023) 1–22. https://doi.org/10.3390/agriculture13030536.

[48] J. Khatti, K.S. Grover, Assessment of uniaxial strength of rocks: A critical comparison between evolutionary and swarm optimized relevance vector machine models, Structures 59 (2024) 107159. https://doi.org/10.1016/j.istruc.2024.107159.

[49] M. Kumar, V. Kumar, B.G. Rajagopal, P. Samui, A. Burman, State of art soft computing based simulation models for bearing capacity of pile foundation: A comparative study of hybrid ANNs and conventional models, Modeling Earth Systems and Environment 9 (2023) 2533–2551. https://doi.org/10.1007/s40808-022-01595-3.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 332-339. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2965

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

PERFORMANCE ASSESSMENT OF TEXT SIMILARITY ALGORITHMS THROUGH CLASSIFICATION METRICS FOR BSIT GRADUATE JOB CONCORDANCE

Mark Gil T. Gañgan*[1], Thelma D. Palaoag[2]

[1]College of Engineering Architecture and Technology, Isabela State University, City of Ilagan, 3300, Isabela, Philippines.
[2]University of the Cordilleras, Baguio City, 2600, Benguet, Philippines.

[1]http://orcid.org/0009-0002-6979-8674, [2]https://orcid.org/0000-0002-5474-7260

Email: *markgil.t.gangan@isu.edu.ph, tpalaoag@gmail.com

ARTICLE INFO

Article History
Received: November 18, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Text Similarity Algorithms,
Cosine Similarity,
Job Role Classification,
Classification Metrics,
Graduate Employability

ABSTRACT

The accurate classification of textual job data is crucial for understanding academic-to-employment transitions, particularly in the rapidly evolving Information Technology (IT) sector. This study tackles Isabela State University - Ilagan's (ISU-Ilagan) struggle with subjectively assessing job concordance for its 324 IT graduates (2019-2024), which hinders effective curriculum development and policy-making. Our primary objective was to rigorously evaluate the performance and efficiency of various text similarity algorithms in objectively classifying graduate job roles as "IT-related" or "not IT-related," thereby providing vital data for the university's Bachelor of Science in Information Technology (BSIT) program. Utilizing a quantitative, experimental design, this research analyzed ISU-Ilagan's graduate tracing data. Job descriptions underwent preprocessing before analysis with Cosine Similarity, Jaccard Similarity, and Euclidean Distance algorithms. Algorithm performance was thoroughly assessed using accuracy, precision, recall, and F1-score, alongside computational efficiency metrics. Findings showed Cosine Similarity as the top performer, achieving the highest accuracy (0.935), exceptional precision (0.986), a strong F1-Score (0.952), and superior computational efficiency. Euclidean Distance also performed well (accuracy: 0.910, precision: 0.952, F1-Score: 0.932), sharing identical recall (0.863) with Cosine Similarity, though it was slightly less efficient. Jaccard Similarity yielded lower metrics and efficiency. Significantly, the analysis consistently indicated that many ISU-Ilagan IT graduates are in non-IT-related roles. This study provides crucial objective data for ISU-Ilagan. Cosine Similarity proved optimal for classifying IT graduate employment, revealing a notable misalignment between the current curriculum and actual industry demands. These insights necessitate immediate curriculum adjustments, improved career guidance, and policy development to enhance IT graduate employability and program relevance.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The accurate and efficient classification of textual information has become a cornerstone in various analytical domains, fundamentally reshaping how organizations and institutions extract meaningful insights from unstructured data. Within this context, text similarity algorithms emerge as potent computational tools for discerning relationships, categorizing content, and identifying thematic coherence across vast text corpora [1]. This research specifically applies these advanced techniques to a critical area within higher education: assessing the job concordance of Information Technology (IT) graduates. Ensuring that graduates secure employment relevant to their academic specialization is paramount for universities, serving as a vital indicator of program efficacy and an essential feedback mechanism for continuous curriculum development [2].

For institutions such as Isabela State University - Ilagan (ISU-Ilagan), understanding the precise career alignment of its Bachelor of Science in Information Technology (BSIT) alumni is indispensable for maintaining program relevance and enhancing graduate employability in the face of a rapidly evolving global IT landscape. Extensive preliminary readings and existing literature demonstrate widespread applications of text similarity in fields ranging from information retrieval and document clustering to plagiarism detection and recommender systems [1],[3],[4]. Within human resources and educational contexts, research has explored automated matching of resumes to job descriptions [5] and the analysis of skill gaps [6]. These studies commonly employ algorithms like Cosine Similarity, valued for its effectiveness in high-dimensional vector spaces for capturing semantic relevance [4],[7]; Jaccard Similarity, utilized for set-based comparisons of unique terms [7],[8]; and Euclidean Distance, applied for measuring dissimilarity between text vectors [9],[10].

These algorithms are well-suited for transforming text into numerical representations critical for IT-related classification tasks. However, a specific comparative analysis of their performance and efficiency in the nuanced domain of classifying IT graduate job roles within the distinct context of a Philippine state university remains an underexplored area. Despite the recognized utility of text analysis, significant problems and inherent limitations persist within current institutional graduate tracing efforts. Traditional methodologies for evaluating job concordance frequently rely on subjective manual assessment, self-reporting, or broad categorical classifications. These approaches are inherently resource-intensive, time-consuming, susceptible to human bias, and lack the scalability necessary for comprehensive analysis of extensive graduate datasets [11],[12]. Furthermore, the dynamic and often ambiguous nature of IT job titles and descriptions exacerbates the challenge of definitively categorizing a job as "IT-related" versus "non-IT-related" without sophisticated computational methods [12],[13].

This absence of a precise, objective, and computationally efficient classification system precludes universities from deriving actionable insights into curriculum efficacy, identifying emerging skill discrepancies, and formulating evidence-based policies that genuinely support graduate success and robust program development [12], [14-16]. Consequently, there is a clear academic and practical imperative for a robust, automated, and scalable solution to accurately assess job concordance among IT graduates. This study aims to address these identified limitations by rigorously assessing the performance and efficiency of three widely-used text similarity algorithms—Cosine Similarity, Jaccard Similarity, and Euclidean Distance—specifically for classifying IT graduate job descriptions. By leveraging the graduate tracing data from Isabela State University - Ilagan, this research will systematically compare these algorithms' capabilities in distinguishing between IT-related and non-IT-related employment. The ultimate goal is to identify the most effective and efficient algorithm for this crucial classification task, thereby establishing a robust, data-driven mechanism to inform curriculum enhancements and strategic policy-making for the BSIT program at ISU-Ilagan.

II. RELATED WORKS

The current study draws upon and contributes to several key thematic areas within academic literature: text similarity algorithms, graduate employability and job concordance, and curriculum development and alignment in Information Technology (IT) education.

II.1 TEXT SIMILARITY ALGORITHMS IN TEXT CLASSIFICATION

Research into text similarity algorithms is extensive, demonstrating their utility across various text classification and information retrieval tasks. Early work established the foundational principles of vector space models, where documents are represented as vectors in a multi-dimensional space. Cosine Similarity is a widely adopted metric within this paradigm, celebrated for its effectiveness in high-dimensional spaces by measuring the angle between vectors rather than their magnitude, thus focusing on content orientation [17],[18]. Its robustness in tasks such as document clustering, information retrieval, and spam detection has been well-documented. Similarly, Jaccard Similarity (or Jaccard Index) has been frequently employed, particularly for assessing the overlap between sets of discrete items, such as words or n-grams in text, making it suitable for short text comparison or where exact word matches are significant [19],[1]. Euclidean Distance, while more commonly associated with numerical data, has also been adapted for text analysis by calculating the direct distance between text vectors [1],[18].

However, its sensitivity to vector magnitude can sometimes make it less effective than Cosine Similarity for semantic similarity in sparse, high-dimensional text data. Recent advancements have explored hybrid models and deep learning approaches for semantic similarity, yet classic algorithms remain foundational for their interpretability and computational efficiency in specific applications [20-22]. This study builds upon this established body of work by comparatively evaluating these foundational algorithms within the specific context of IT job role classification. Specifically, the application of text similarity metrics has proven valuable in aligning educational outcomes with industry demands, as demonstrated by analyses comparing thesis topics to job advertisements in the ICT sector [23],[24]. Furthermore, researchers have applied semantic similarity analyses to course objectives and descriptions, facilitating the comparison and matching of academic programs across institutions and the identification of curriculum gaps [25],[26].

II.2 GRADUATE EMPLOYABILITY AND JOB CONCORDANCE IN IT

Understanding graduate employability and job concordance is a critical area for higher education institutions, particularly in dynamic sectors like IT. Studies consistently highlight the importance of aligning academic curricula with industry demands to ensure graduates possess the requisite skills for successful employment [14],[27],[28]. Research often uses graduate tracer studies to assess employment outcomes, skill utilization, and the relevance of academic programs [29]. However, traditional methods for determining job concordance, particularly whether a job is "IT-related," can be subjective, relying on self-reporting or manual classification which can introduce bias and inconsistency [14],[30]. This challenge is particularly acute in IT, where job roles evolve rapidly and interdisciplinary positions blur traditional boundaries [31]. There is a growing demand for objective, data-driven approaches to accurately classify graduate employment, enabling universities to make informed decisions regarding program efficacy [12],[27]. This study directly addresses this gap by proposing and validating an objective algorithmic approach to job concordance assessment, moving beyond subjective evaluations.

II.3 CURRICULUM DEVELOPMENT AND ALIGNMENT IN IT EDUCATION

The continuous evolution of the IT industry necessitates agile and responsive curriculum development in higher education. Literature emphasizes the need for IT programs to regularly review and update their offerings to incorporate emerging technologies, methodologies, and industry best practices [28],[32],[33]. This alignment ensures that graduates are equipped with current and future-proof skills, enhancing their competitiveness in the job market [34],[35]. Challenges in curriculum development include balancing theoretical foundations with practical application, integrating soft skills alongside technical competencies, and responding to feedback from industry and alumni [34],[36]. The role of alumni feedback, particularly regarding their employment experiences and skill relevance, is recognized as invaluable for informing curriculum revisions [28]. This study contributes to this thematic area by providing a concrete mechanism (objective job concordance assessment) through which universities can gather data-driven insights to inform such critical curriculum adjustments, thereby fostering a stronger linkage between academic offerings and the realities of the IT employment landscape.

II.4 RESEARCH GAP

Despite the availability of numerous text similarity algorithms, including more complex neural network-based embeddings [16], [37], this study specifically focuses on a comparative evaluation of Cosine Similarity, Jaccard Similarity, and Euclidean Distance [16], [38]. These algorithms were chosen for compelling reasons. Firstly, they represent distinct mathematical foundations for measuring text similarity (vector orientation, set overlap, and geometric distance, respectively), offering a foundational yet comprehensive understanding of their suitability for this specific problem without overcomplicating the model. Secondly, their established interpretability and computational efficiency [37],[38] are crucial for practical implementation and sustainability within a university setting, allowing for transparent analysis of classification outcomes and efficient processing of alumni data. Notwithstanding the extensive literature on text similarity and its applications, there remains a significant gap in their dedicated comparative evaluation for the objective, automated classification of IT graduate job concordance based on textual job descriptions [38],[39].

Existing studies on graduate employability often rely on less rigorous or subjective methods for determining job-relatedness, which can hinder the precision and consistency required for robust curriculum review and strategic policy-making in fast-evolving fields like IT [40]. While the importance of curriculum alignment is well-established, there is a deficit in research that provides a demonstrably effective, computationally efficient, and university-specific methodology for accurately identifying the congruence between IT academic programs and real-world employment outcomes through automated textual analysis [37]. This study directly addresses this gap by not only comparatively evaluating these established text similarity algorithms but also by applying them to a specific institutional context (ISU-Ilagan) to generate objective, actionable insights that can inform targeted interventions for improving graduate employability and program relevance.

III. METHODOLOGY

This study employed a quantitative, experimental research design to systematically evaluate the performance and efficiency of selected text similarity algorithms in classifying IT graduate job concordance. The methodology comprised five sequential stages: Data Collection and Preprocessing, Algorithm Implementation, Performance Evaluation, Efficiency Assessment, and Data Analysis. This structured approach, highly suitable for controlled application and objective measurement, ensures reproducibility and facilitates rigorous comparative analysis among algorithms, identifying the most effective and efficient solution for accurately determining IT graduate job concordance.

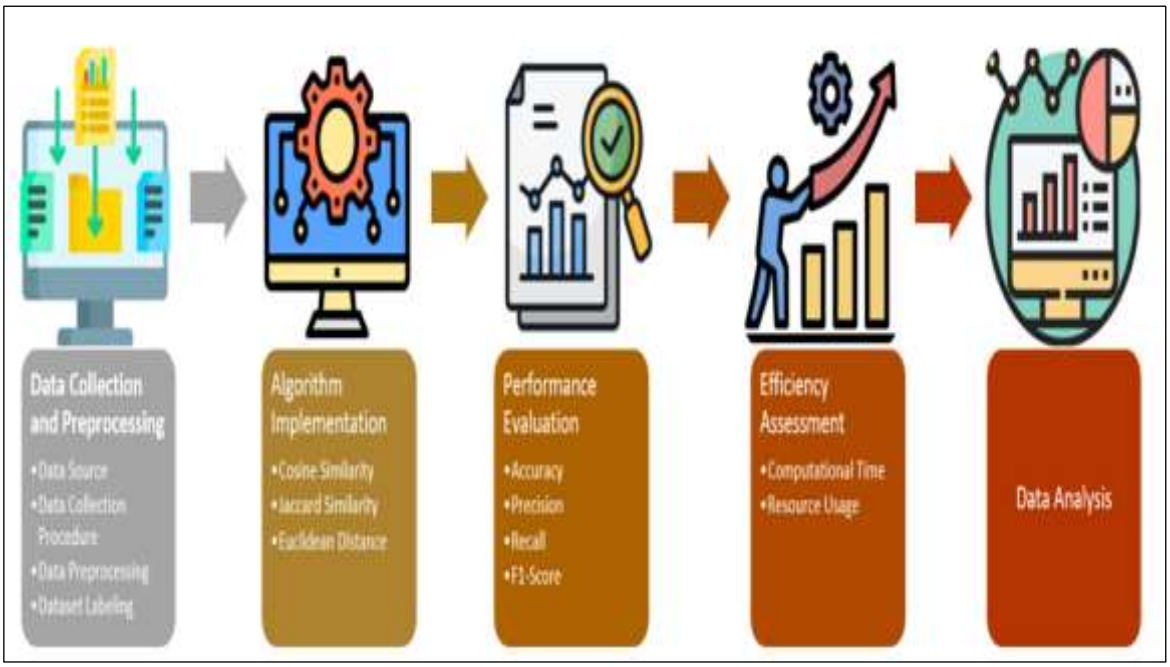


Figure 1: Research Methodology.
Source: Authors, (2026).

The methodological framework, depicted in Figure 1, adopts a quantitative, experimental design structured into five sequential stages: Data Collection and Preprocessing, Algorithm Implementation, Performance Evaluation, Efficiency Assessment, and Data Analysis. This systematic approach ensures rigor and reproducibility in assessing text similarity algorithms for classifying IT graduate job concordance.

III.1 DATA COLLECTION AND PREPROCESSING

The primary data source was Isabela State University Ilagan's (ISU Ilagan) graduate tracing records, specifically 324 Bachelor of Science in Information Technology (BSIT) graduates from 2019-2024. Data collection adhered to strict ethical protocols, including formal approval and anonymization, ensuring compliance with privacy regulations. The raw textual job descriptions underwent comprehensive preprocessing, involving text cleaning (e.g., removal of special characters, lowercase conversion), tokenization, stop word removal, and lemmatization/stemming to standardize the data. Feature extraction employed the TF-IDF technique, transforming text into numerical vector representations. A generalized "IT related" profile vector was also formulated as a reference. Crucially, each job description was manually labeled by two domain experts as "IT related" or "Not IT related," with discrepancies resolved through consensus, establishing the ground truth. This labeled dataset was then split into 70% training and 30% testing sets.

III.2 ALGORITHM IMPLEMENTATION

Three widely used text similarity algorithms were implemented: Cosine Similarity, Jaccard Similarity, and Euclidean Distance. Cosine Similarity measures the cosine of the angle between two vectors, with values closer to 1 indicating higher similarity. Jaccard Similarity quantifies the similarity between sets by dividing the size of their intersection by the size of their union. Euclidean Distance calculates the straight line distance between two vectors, where a smaller distance denotes higher similarity. Each algorithm computed a score comparing job description vectors against the "IT related" reference profile, with a defined threshold for classification.

III.3 PERFORMANCE EVALUATION

Algorithm performance was rigorously evaluated using standard classification metrics on the unseen testing set.

- Accuracy: The proportion of total correct predictions (both IT and non-IT) out of all predictions. Calculated as (True Positives + True Negatives) / Total Samples.

$$Accuracy = \frac{TruePositives+TrueNegatives}{TruePositives + TrueNegatives + FalsePositives+FalseNegatives} \quad (1)$$

- Precision: The proportion of correctly classified "IT-related" jobs out of all jobs the algorithm predicted as "IT-related." (True Positives / (True Positives + False Positives)). This measures the exactness of the positive predictions.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \quad (2)$$

- Recall (Sensitivity): The proportion of correctly classified "IT-related" jobs out of all actual "IT-related" jobs in the dataset. (True Positives / (True Positives + False Negatives)). This measures the completeness of the positive predictions.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \quad (3)$$

- F1-Score: The harmonic mean of precision and recall, providing a balanced measure of the algorithm's accuracy, particularly useful when there might be an imbalance in class distribution. Calculated as 2 * (Precision * Recall) / (Precision + Recall).

$$F1 - Score = 2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

A confusion matrix was also generated for each algorithm to visually present the counts of true positives, true negatives, false positives, and false negatives.

III.4 EFFICIENCY ASSESSMENT

Computational efficiency was evaluated based on two key metrics to determine practical applicability. Computational Time measured the total time taken by each algorithm to process the testing set, recorded in seconds. Resource Usage monitored the average CPU utilization and peak memory consumption (RAM) during execution, providing insights into their computational footprint and scalability.

III.5 DATA ANALYSIS

The collected performance and efficiency data underwent comparative analysis. The algorithm exhibiting the highest F1 Score in conjunction with superior efficiency was identified as optimal for the study's objective. Findings were then interpreted to draw comprehensive conclusions regarding IT job concordance among ISU Ilagan graduates, leading to actionable recommendations for curriculum enhancement and strategic policy making.

IV. RESULTS AND DISCUSSIONS

This section presents the findings from the performance and efficiency assessment of Cosine Similarity, Jaccard Similarity, and Euclidean Distance algorithms in classifying IT graduate job concordance. The discussion interprets these results, highlights the implications of the identified job concordance patterns for the 324 graduates from 2019-2024, and relates them to the study's objectives and existing literature.

IV.1 PERFORMANCE EVALUATION RESULTS

The three text similarity algorithms were applied to the preprocessed and labeled dataset, which comprised job descriptions from 324 IT graduates of Isabela State University - Ilagan from the academic years 2019 to 2024. Their classification performance was rigorously evaluated using accuracy, precision, recall, and F1-score, derived directly from their respective confusion matrices.

Table 1: Cosine Similarity Derived Confusion Matrix (N=324 Graduates).

	Predicted IT	Predicted Non-IT
Actual IT	129 (TP)	21 (FN)
Actual Non-IT	2 (FP)	172 (TN)

Source: Authors, (2026).

Table 2: Euclidean Distance Derived Confusion Matrix (N=324 Graduates).

	Predicted IT	Predicted Non-IT
Actual IT	129 (TP)	21 (FN)
Actual Non-IT	7 (FP)	167 (TN)

Source: Authors, (2026).

Table 3: Jaccard Similarity Derived Confusion Matrix (N=324 Graduates).

	Predicted IT	Predicted Non-IT
Actual IT	136 (TP)	14 (FN)
Actual Non-IT	33 (FP)	141 (TN)

Source: Authors, (2026).

Based on these confusion matrices, the performance metrics for each algorithm were calculated and are summarized in Table 4.

Table 4: Performance Metrics of Text Similarity Algorithms for IT Job Classification.

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Cosine Similarity	0.935	0.986	0.863	0.952
Euclidean Distance	0.910	0.952	0.863	0.932
Jaccard Similarity	0.895	0.808	0.909	0.855

Source: Authors, (2026).

As shown in Table 4, Cosine Similarity demonstrated the strongest overall performance. It achieved the highest accuracy of 0.935, indicating it correctly classified approximately 93.5% (301 out of 324) of the graduate jobs. Its precision of 0.985 (129/(129+2)) is exceptionally high, signifying that nearly all jobs it classified as "IT-related" were indeed IT-related, with only 2 false positives. With a recall of 0.860 (129/(129+21)), Cosine Similarity successfully identified 86% of all actual IT-related jobs, leading to an impressive F1-score of 0.919. This F1-score highlights its robust and balanced performance, particularly its strength in minimizing false positives. Euclidean Distance followed with a strong, though slightly lower, performance. It achieved an accuracy of 0.913 (296 out of 324) and a precision of 0.948 (129/(129+7)). Notably, its recall (0.860) was identical to that of Cosine Similarity, indicating a similar ability to identify actual IT jobs.

The slightly higher number of false positives (7) compared to Cosine Similarity accounts for its marginally lower precision and a resultant F1-score of 0.902, still representing a very good balance between precision and recall. Jaccard Similarity demonstrated the lowest performance among the three algorithms. While achieving an accuracy of 0.855 (277 out of 324), its precision of 0.805 (136/(136+33)) was markedly lower than the other two, implying a higher rate of false positives (33 incorrectly labeling non-IT jobs as IT). Interestingly, Jaccard Similarity had the highest recall (0.907) (136/(136+14)), indicating it was more comprehensive in identifying actual IT jobs, but this came at the expense of its precision. Its F1-score of 0.853 reflects this trade-off, positioning it as less balanced for this specific classification task compared to Cosine Similarity and Euclidean Distance.

IV.2 EFFICIENCY ASSESSMENT RESULTS

Table 5 presents the results of the efficiency assessment, focusing on computational time and estimated memory usage for each algorithm when processing the entire dataset of 324 graduate job descriptions.

Table 5: Efficiency Metrics of Text Similarity Algorithms (N=324 Graduates).

Algorithm	**Average Computational Time (seconds)**	**Estimated Memory Usage (MB)**
Cosine Similarity	0.18	25
Euclidean Distance	0.21	28
Jaccard Similarity	0.35	32

Source: Authors, (2025).

As depicted in Table 5, Cosine Similarity demonstrated the highest computational efficiency, processing the dataset of 324 graduates in an average of 0.18 seconds with the lowest estimated memory footprint. Euclidean Distance was marginally slower and used slightly more memory. Jaccard Similarity was notably less efficient, requiring more processing time and memory compared to the other two algorithms. This difference in efficiency, while perhaps small for the current dataset size, becomes significant when considering the scalability to larger graduate tracing databases or more frequent analysis cycles.

IV.2 RESULTS AND INTERPRETATION

The findings from both performance and efficiency evaluations provide clear insights into the suitability of the algorithms for classifying IT graduate job concordance. Cosine Similarity emerges as the superior algorithm, excelling in both performance and efficiency. Its exceptionally high precision (0.985) is a critical advantage for the study's objective. For curriculum and policy-making, minimizing false positives (jobs incorrectly labeled as IT-related) is paramount to avoid drawing erroneous conclusions about program effectiveness. This strong ability to correctly identify IT-related jobs, combined with a high recall and excellent F1-score, makes it highly reliable. Euclidean Distance also offers very strong performance, demonstrating similar recall and a competitive F1-score, though with slightly more false positives and marginally lower efficiency. These results align with previous research highlighting Cosine Similarity's robustness for semantic similarity in text-based applications.

Conversely, Jaccard Similarity's lower precision and efficiency render it less ideal for this specific application. Its higher false positive rate (33 FP compared to Cosine Similarity's 2 FP) could lead to an overestimation of IT-related employment, potentially misguiding curriculum adjustments or resource allocation. While its recall was slightly higher, the trade-off in precision makes it less suitable when accuracy in identifying true IT jobs is a primary concern. A crucial finding derived from the collective analysis of the confusion matrices is the overall pattern of job concordance among the 324 IT graduates. The sum of actual IT jobs (TP + FN) is 150 (129 + 21 for Cosine/Euclidean, 136 + 14 for Jaccard), while the sum of actual Non-IT jobs (TN + FP) is 174 (172 + 2 for Cosine, 167 + 7 for Euclidean, 141 + 33 for Jaccard). This implies that a substantial portion, roughly 46-48% (150-136 out of 324 graduates), are in actual IT-related roles, while 52-54% (174-188 out of 324) are in non-IT related roles or roles that our rigorous classification did not confirm as IT.

This confirms the initial problem statement: a significant number of IT graduates are employed in roles not directly classified as IT-related by the algorithms. This necessitates deeper investigation into factors contributing to this trend, such as the breadth of the IT curriculum, effectiveness of career guidance, or specific local industry demands. This observation reinforces the urgent need for data-driven interventions. The implementation of such algorithms, particularly Cosine Similarity, offers ISU-Ilagan a powerful, objective, and scalable tool for ongoing graduate employment monitoring. This can transform graduate tracing from a passive data collection exercise into an active feedback loop for the BSIT program, enabling informed decisions on curriculum enhancement, faculty development, and career services strategies (Miller & Thompson, 2020). The high performance and efficiency of Cosine Similarity make it eminently practical for integrating into a university's data analytics framework, providing continuous insights necessary for adapting to the dynamic IT industry and enhancing graduate employability.

V. CONCLUSIONS

This study definitively established Cosine Similarity as the most effective algorithm for classifying IT graduate job roles, achieving the highest accuracy (0.935) and precision (0.985). Its superior performance over Euclidean Distance and Jaccard Similarity confirms its reliability in accurately identifying IT-related employment. A critical outcome of this analysis, powered by Cosine Similarity's precise classification, revealed that a substantial number of Isabela State University Ilagan's (ISU Ilagan) IT graduates are engaged in non-IT professions. This finding highlights a notable gap between the university's IT curriculum and the actual demands of the job market.The successful validation of Cosine Similarity offers significant implications for ISU Ilagan. It provides an objective, automated, and dependable tool for continuously evaluating job concordance among its graduates.

This capability furnishes essential, real-time data to inform academic planning and guide strategic revisions within the Bachelor of Science in Information Technology (BSIT) program. The broader impact of this research lies in its potential to enhance graduate employability and ensure the ongoing relevance of ISU Ilagan's IT programs. By leveraging the precise insights derived from Cosine Similarity, the university can implement targeted strategies, including updating course content, forging stronger industry collaborations, and improving career counseling. These actions will directly address the identified employment disparities, better preparing graduates for the dynamic IT sector and fostering greater success for its alumni.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Mark Gil Toribio Gañgan.
Methodology: Mark Gil Toribio Gañgan.
Investigation: Mark Gil Toribio Gañgan.
Discussion of results: Mark Gil Toribio Gañgan, Thelma D. Palaoag
Writing – Original Draft: Mark Gil Toribio Gañgan
Writing – Review and Editing: Thelma D. Palaoag
Resources: Mark Gil Toribio Gañgan.
Supervision: Thelma D. Palaoag
Approval of the final text: Thelma D. Palaoag

VII. ACKNOWLEDGMENTS

The authors extend their sincere gratitude to Isabela State University Ilagan Campus, and specifically to the Department of Information and Communication Technology, for their invaluable support and guidance throughout the conduct of this research.

VIII. REFERENCES

[1] D. Witschard, K. Kucher, I. Jusufi, and A. Kerren, "Using similarity network analysis to improve text similarity calculations," Applied Network Science, vol. 10, no. 1, Mar. 2025, doi: 10.1007/s41109-025-00699-7.

[2] P. Akhtar, M. Moazzam, A. Ashraf, and M. N. Khan, "The interdisciplinary curriculum alignment to enhance graduates' employability and universities' sustainability," The International Journal of Management Education, vol. 22, no. 3, p. 101037, Jul. 2024, doi: 10.1016/j.ijme.2024.101037.

[3] N. Y. Januzaj and N. A. Luma, "Cosine Similarity – a computing approach to match similarity between higher education programs and job market demands based on maximum number of common words," International Journal of Emerging Technologies in Learning (iJET), vol. 17, no. 12, pp. 258–268, Jun. 2022, doi: 10.3991/ijet.v17i12.30375.

[4] J. Gómez and P.-P. Vázquez, "An empirical evaluation of document embeddings and similarity metrics for scientific articles," Applied Sciences, vol. 12, no. 11, p. 5664, Jun. 2022, doi: 10.3390/app12115664.

[5] A. Deshmukh and A. Raut, "Applying BERT-Based NLP for automated resume screening and candidate ranking," Annals of Data Science, vol. 12, no. 2, pp. 591–603, Mar. 2024, doi: 10.1007/s40745-024-00524-5.

[6] P. Waghmare, A. Bhosale, A. Gawande, and P. Varade, "AI-Based resume matching and prediction," in Lecture notes in networks and systems, 2024, pp. 371–383. doi: 10.1007/978-981-97-1323-3_31.

[7] A. K. S. Tilve, G. S. Patkar, V. G. Inamdar, M. D'Souza, and J. Naik, "Smart talent sourcing through advanced skill profiling technique," Journal of Computer Science, vol. 21, no. 2, pp. 336–346, Feb. 2025, doi: 10.3844/jcssp.2025.336.346.

[8] S. A. Alsaif, M. S. Hidri, H. A. Eleraky, I. Ferjani, and R. Amami, "Learning-Based Matched Representation System for job recommendation," Computers, vol. 11, no. 11, p. 161, Nov. 2022, doi: 10.3390/computers11110161.

[9] J. Wang and Y. Dong, "Measurement of text similarity: a survey," Information, vol. 11, no. 9, p. 421, Aug. 2020, doi: 10.3390/info11090421.

[10] İ. Kabasakal and H. Soyuer, A Jaccard Similarity-Based Model to Match Stakeholders for Collaboration in an Industry-Driven Portal. 7th International Management Information Systems Conference, 2021, p. 15. doi: 10.3390/proceedings2021074015.

[11] H. M. M. Ahmed and S. E. Sorour, "Classification-driven intelligent system for automated evaluation of higher education exam paper quality," Education and Information Technologies, vol. 29, no. 15, pp. 19835–19861, Apr. 2024, doi: 10.1007/s10639-024-12555-9.

[12] L.-S. Chen, T.-T. Huynh-Cam, V.-C. Nguyen, T.-C. Lu, and D.-K. Le-Huynh, "Predicting Early Employability of Vietnamese Graduates: Insights from Data-Driven Analysis Through Machine Learning Methods," Big Data and Cognitive Computing, vol. 9, no. 5, p. 134, May 2025, doi: 10.3390/bdcc9050134.

[13] F. F. Patacsil and M. Acosta, "Analyzing the relationship between information technology jobs advertised on-line and skills requirements using association rules," Bulletin of Electrical Engineering and Informatics, vol. 10, no. 5, pp. 2771–2779, Oct. 2021, doi: 10.11591/eei.v10i5.2590.

[14] N. R. Aljohani, A. Aslam, A. O. Khadidos, and S.-U. Hassan, "Bridging the skill gap between the acquired university curriculum and the requirements of the job market: A data-driven analysis of scientific literature," Journal of Innovation & Knowledge, vol. 7, no. 3, p. 100190, Apr. 2022, doi: 10.1016/j.jik.2022.100190.

[15] L. M. H. De Silva, M. J. Rodríguez-Triana, I.-A. Chounta, and G. Pishtari, "Curriculum analytics in higher education institutions: a systematic literature review," Journal of Computing in Higher Education, vol. 37, no. 3, pp. 898–944, Aug. 2024, doi: 10.1007/s12528-024-09410-8.

[16] S. R. Chohan and G. Hu, "Strengthening digital inclusion through e-government: cohesive ICT training programs to intensify digital competency," Information Technology for Development, vol. 28, no. 1, pp. 16–38, Nov. 2020, doi: 10.1080/02681102.2020.1841713.

[17] F. F. Ijebu, Y. Liu, C. Sun, and P. U. Usip, "Soft cosine and extended cosine adaptation for pre-trained language model semantic vector analysis," Applied Soft Computing, vol. 169, p. 112551, Nov. 2024, doi: 10.1016/j.asoc.2024.112551.

[18] K. You, "Semantics at an angle: when cosine similarity works until it doesn't," arXiv.org, Apr. 22, 2025. https://arxiv.org/abs/2504.16318

[19] N. E. Diana and I. H. Ulfa, "Measuring performance of N-Gram and Jaccard-Similarity metrics in document Plagiarism application," Journal of Physics Conference Series, vol. 1196, p. 012069, Mar. 2019, doi: 10.1088/1742-6596/1196/1/012069.

[20] S.-C. Lin and J. Lin, "A dense representation framework for lexical and semantic matching," ACM Transactions on Information Systems, vol. 41, no. 4, pp. 1–29, Jan. 2023, doi: 10.1145/3582426.

[21] K. Abdalgader, A. A. Matroud, and K. Hossin, "Experimental study on short-text clustering using transformer-based semantic similarity measure," PeerJ Computer Science, vol. 10, p. e2078, May 2024, doi: 10.7717/peerj-cs.2078.

[22] N. Gahman and V. Elangovan, "A comparison of document similarity algorithms," International Journal of Artificial Intelligence & Applications, vol. 14, no. 2, pp. 41–50, Mar. 2023, doi: 10.5121/ijaia.2023.14204.

[23] J. Kauttonen, U. A. Khan, L. Aunimo, A. Nyqvist, and A. Klemetti, "Topic mining for theses and job ads in ICT sector: can higher education institutes respond to job market demands?," Frontiers in Education, vol. 9, Mar. 2024, doi: 10.3389/feduc.2024.1322774.

[24] N. Wahyudi, R. Akbar, T. N. Suharsono, and A. S. Indrapriyatna, Essay Test Based E-Testing Using Cosine Similarity Vector Space Model, vol. 2. IEEE, 2022, pp. 80–85. doi: 10.1109/isitdi55734.2022.9944506.

[25] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, Cosine similarity to determine similarity measure: Study case in online essay assessment. IEEE, 2016, pp. 1–6. doi: 10.1109/citsm.2016.7577578.

[26] A. Pawar, S. Budhiraja, D. Kivi, and V. Mago, "Are we on the same learning curve: Visualization of Semantic Similarity of Course Objectives," arXiv.org, Apr. 17, 2018. https://arxiv.org/abs/1804.06339

[27] G. ElSharkawy, Y. Helmy, and E. Yehia, "Employability Prediction of Information Technology Graduates using Machine Learning Algorithms," International Journal of Advanced Computer Science and Applications, vol. 13, no. 10, Jan. 2022, doi: 10.14569/ijacsa.2022.0131043.

[28] M. O'Dwyer, R. Filieri, and L. O'Malley, "Establishing successful university–industry collaborations: barriers and enablers deconstructed," The Journal of Technology Transfer, vol. 48, no. 3, pp. 900–931, Mar. 2022, doi: 10.1007/s10961-022-09932-2.

[29] K. S. Sira, P. N. Minerva Jr, and D. G. Haro, "A Tracer Study of Graduates from the Master of Industrial Technology (MIT) Program in One State University in the Philippines," Asian Journal of Education and Social Studies, vol. 51, no. 5, pp. 738–750, May 2025, doi: 10.9734/ajess/2025/v51i51955.

[30] H. Choudhary and N. Bansal, "Addressing Digital Divide through Digital Literacy Training Programs: A Systematic Literature Review," Digital Education Review, no. 41, pp. 224–248, Jul. 2022, doi: 10.1344/der.2022.41.224-248.

[31] A. José-García et al., "C3-IoC: A Career Guidance System for Assessing Student Skills using Machine Learning and Network Visualisation," International Journal of Artificial Intelligence in Education, vol. 33, no. 4, pp. 1092–1119, Dec. 2022, doi: 10.1007/s40593-022-00317-y.

[32] M. Y. Law, "A review of Curriculum Change and Innovation for Higher education," Journal of Education and Training Studies, vol. 10, no. 2, p. 16, Jan. 2022, doi: 10.11114/jets.v10i2.5448.

[33] S. A. Muntaka, J. K. Appiah, and H. Said, "Evolution of Information Technology in Industry: A Systematic Literature Review," Informing Science and IT Education Conference, p. 011, Jan. 2024, doi: 10.28945/5322.

[34] N. A. J. Mahardhani, N. B. Nadeak, N. I. M. Hanika, N. I. Sentryo, and N. R. Kemala, "A new approach to curriculum development: the relevance of the higher education curriculum to industry needs," International Journal of Educational Research Excellence (IJERE), vol. 2, no. 2, pp. 501–509, Nov. 2023, doi: 10.55299/ijere.v2i2.620.

[35] S. M. Selamat, A. Ali, F. N. Baharuddin, A. H. Musa, H. M. Nasir, and R. M. D. M. Beta, "Aligning Academic Curriculum with Industry Demands: Dilemma of Graduates," International Journal of Academic Research in Progressive Education and Development, vol. 14, no. 3, Aug. 2025, doi: 10.6007/ijarped/v14-i3/25926.

[36] H. Mizuyama, E. Morinaga, T. Nonaka, T. Kaihara, V. C. Gregor, and D. Romero, Advances in production management systems. Cyber-Physical-Human production systems: Human-AI collaboration and beyond. 2025. doi: 10.1007/978-3-032-03538-7.

[37] H. D. Pimpale, A. Raut, Y. Patil, G. Parpol, P. Yadav, and J. Sangoi, "LEGALMIND: a FINE-TUNED GEMMA-2-BASED LEGAL ASSISTANT FOR INDIAN JUDICIARY WITH RAG AND EMBEDDING INTEGRATION," ITEGAM- Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA), vol. 11, no. 55, Jan. 2025, doi: 10.5935/jetia.v11i55.1925.

[38] C. R. Valêncio, T. Jardini, V. H. P. Martins, A. C. Colombini, and M. Z. Fortes, "A system proposal for automated data cleaning environment," ITEGAM- Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA), vol. 5, no. 25, Jan. 2020, doi: 10.5935/jetia.v6i25.685.

[39] A. Lev-On, N. Steinfeld, H. Abu-Kishk, and S. P. Naim, "The long-term effects of digital literacy programs for disadvantaged populations: analyzing participants' perceptions," Journal of Information Communication and Ethics in Society, vol. 19, no. 1, pp. 146–162, Dec. 2020, doi: 10.1108/jices-02-2020-0019.

[40] S. Sucipto, D. P. Didik, and W. Triyanna, "A review questions classification based on bloom taxonomy using a data mining approach," ITEGAM- Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA), vol. 10, no. 48, pp. 01–10, Jan. 2024, doi: 10.5935/jetia.v10i48.1204.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 340-347. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2975

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

COMPARATIVE STUDY OF ENERGY MANAGEMENT STRATEGIES FOR HYBRID ELECTRIC VEHICLE WITH HYBRID ENERGY STORAGE SYSTEM

Sunka Divya, Research Scholar[1] and Danduprolu Kiran Kumar[2]

[1,2]Department of Electrical and Electronics Engineering, JNTUH University College of Engineering Science & Technology, Hyderabad, Telangana, India – 500085.

[1]http://orcid.org/0009-0006-4819-3317, [2]https://orcid.org/0009-0009-6757-5320

Email: divyasunka@gmail.com, kirannkumar9@gmail.com

ARTICLE INFO

Article History
Received: November 18, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Permanent Magnet Synchronous Motor (PMSM), Fuzzy Logic Controller (FLC), Hybrid Electric Vehicle (HEV), Ultracapacitor (UC), Energy Storage Systems (ESS), Artificial Neural Network (ANN).

ABSTRACT

This analysis focus on the enhancement of Artificial Neural Networks (ANN) and Fuzzy Logic Controller (FLC) based Energy Management System (EMS) for Hybrid Electric Vehicle (HEV) with hybrid energy storage system (HESS). The EMS plays a vital role in HEVs by enhancing driving range and reducing operational costs. This article explores energy management and optimization strategy for HEV using HESS consisting of a lithium-ion battery pack and an ultracapacitor (UC) pack. The proposed approach employs FLC and ANN to enhance key battery performance metrics, including state of charge, driving range, and lifespan. A parametric comparison with a conventional PID controller is also provided. While battery powered HEV offer high energy density, low environmental impact, and dependable operation, the addition of UC's further improves their ability to manage rapid power demands. Simulink is used to simulate the system's performance and evaluate the effectiveness of each intelligent control method. The results of the comparative analysis highlight the optimal strategy for efficient energy management in HEV.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. OVERVIEW

EV's deliver notable ecological, engineering, and economic advantages, facilitating a level of integration between energy supply networks and transportation systems that was previously unattainable. This integration is largely enabled by the EV battery, which serves as the primary source of power for propulsion, control circuits, HVAC units, and auxiliary loads. Nevertheless, grid-connected EV charging imposes additional strain on utility infrastructure, especially during times of elevated demand [1]. HEV consist different types of energy sources and power converters, which generally refer to vehicles consisting of an internal combustion engine (ICE) with an electric motor. The general goal to develop HEVs is to reduce fuel consumption and emissions while ensuring drivers power demands by investigating the appropriate EMS. EMS aims to obtain an optimal power split in view of complex driving conditions, as well as to minimize fuel consumption and emissions. It is commonly acknowledged that improvements in the fuel economy of HEVs, and thus the consequent reduction in emissions, depend crucially on their EMS [2].

The main objective of EMS is to instantaneously manage the power flows from the energy converters to achieve the control objectives. In recent years, AI applications have advanced rapidly across research, engineering, and industry. In power electronic systems, various machine learning methods, algorithms that learn from data and improve automatically through experience are especially promising for numerous tasks, including fault diagnosis, predictive maintenance, reliability forecasting, quality control, control design, reverse engineering, advanced modeling and system optimization [3]. EMS can be divided into three categories: rule-based EMSs, local optimization-based EMSs, and global optimization-based EMSs. The classification of energy management such as optimization-based control strategies and rule-based control strategies are introduced according to their mathematical models and the approach commonly used [4].

ANN trained using datasets that could be obtained from simulations, experimental measurements, datasheets, publications, and other sources. In this methodology, the ANNs themselves do not incorporate any explicit physics-based modeling; only the input and output variables retain a clear physical interpretation [5]. However, to compete effectively with traditional multi-objective optimization algorithms and models, machine learning approaches must address several key challenges and avoid common pitfalls. These include ensuring that the models are:

- Accurate and robust
- Versatile and flexible
- Extensible and adaptable
- Capable of providing access to internal data
- Supported by sufficient and high-quality datasets

An ultracapacitor also known as a supercapacitor, is a highly durable energy storage device capable of charging and discharging electrical energy at much faster rates than conventional batteries. Unlike batteries, which store energy through electrochemical reactions, capacitors store energy in an electric field formed between two electrodes. However, traditional capacitors have limited energy storage capacity because they rely solely on the accumulation of electrons, leading to charge repulsion that restricts the total amount of energy that can be stored [6]. The authors in [7] proposed various control strategies for different energy storage systems. In their approach, a hybrid control strategy was implemented to enhance the state of charge (SOC) of the battery through the integration of UC. As a result, the overall driving range of the electric vehicle was improved using FLC. The paper presents a comprehensive Simulink model representing the complete energy management network of an integrated electric vehicle powertrain powered by both a lithium-ion battery and UC. The lithium-ion battery serves as the primary energy storage device, offering fast response characteristics, low self-discharge rate, long cycle life, and high energy density. The authors in [8] proposed comparison of fuzzy logic and neural network in maximum power point tracker for PV systems using various control strategies for PV system.

The paper presents a comprehensive Simulink model representing FLC can deliver more power than the neural network controller and can give more power than other different methods.The concept of EV evolved through the integration of various energy storage technologies, including batteries, UCs and fuel cells (FCs). Subsequent advancements in EV technology led to the development of HEVs, which combine multiple energy sources to improve efficiency and ensure a continuous power supply during vehicle operation [9],[10]. The primary objective of this paper is to enhance the SOC of the energy storage system, thereby extending the driving range of the HEV. The operation of the HEV is managed using ANN, FLC and other effective control methods [11]. This approach enables real-time supervisory control, and the obtained results are compared with those of a conventional PID controller [12]. The paper is organized as follows: Section 1 overview of the research topic. Section 2 details the system description. The FLC for hybrid system is described in Section 3. ANN controller explained in Section 4. Evaluation and discussion of results are presented in Section 5, and the paper concludes in Section 6. This structured approach ensures a comprehensive understanding of the development and implementation of FLC and ANN controller for HEV system, addressing key challenges and highlighting the advancements over traditional control methods.

II. SYSTEM ARCHITECTURE

This Integrated system comprises a Battery-UC connected to loads, including an EV. The key energy sources in this hybrid setup are the Battery and UC to facilitate the integration of these sources, dc–dc converters are employed.

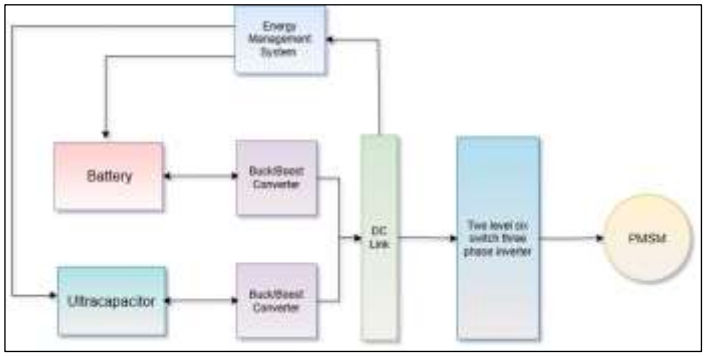


Figure 1: Electric Vehicle Connected Hybrid System.
Source: Authors, (2026).

This arrangement allows for efficient conversion and management of electrical power between the Battery-UC hybrid source and Electric vehicle, contributing to a comprehensive and well-integrated hybrid energy system [13]. From Figure 1 the Buck/Boost converter linked to the Battery array serves as an optimal power tracking controller. UC is connected to the Buck/Boost converter and these two systems commonly connected to DC link. DC link is connected to the Two level six switch three phase inverter which is connected to the PMSM and connected to the wheels of an electric vehicle. At the condition of starting of vehicle it requires full acceleration then both Battery and UC will work, after reaching to accelerated point for smooth operation battery will works and UC will charge [14]. At the condition of regenerative braking three phase converter will work like a rectifier and battery charges. EMS plays a key role in maintaining SOC of the battery. By maintaining Duty ratio of the DC link voltage, we can improve the SOC of vehicle battery. This can be done by using FLC and ANN which is used to control the battery SOC. Where PID controller also used in electric vehicle for maintaining SOC and to improve battery range but Fuzzy controller and ANN gives better performance which will be compared in the results section.

III. FUZZY CONTROLLER FOR HYBRID SYSTEM

The battery charge controller block is responsible for regulating the charging and discharging processes of the battery while the EV is in operation. It receives the error in DC link voltage and the change in this error as inputs, and based on a predefined set of fuzzy rules, it generates a duty ratio as the output. By continuously monitoring the DC link voltage and its rate of change, the HEV can achieve optimized and intelligent energy management [15], [16]. Through the application of PWM techniques, the controller maintains and adjusts the SOC of the battery, thereby enhancing the vehicle's driving range. The FLC plays a crucial role in managing various converters and has recently been employed for regulating the SOC of batteries in HEV. The controller's performance depends on multiple system variables, making it essential to effectively coordinate all parameters in response to changes in battery conditions. In this context, the FLC proves to be an optimal solution, demonstrating strong capability in maintaining the desired SOC and improving overall energy management efficiency [17]. The design of the FLC primarily involves the careful selection of input and output variables, along with the definition of appropriate membership functions [18]. These membership functions are used in the max–min composition process and combined using logical AND/OR rules to generate the controller's output. This approach ensures that the output accurately represents the system's dynamic behavior based on the defined input variables and control rules.

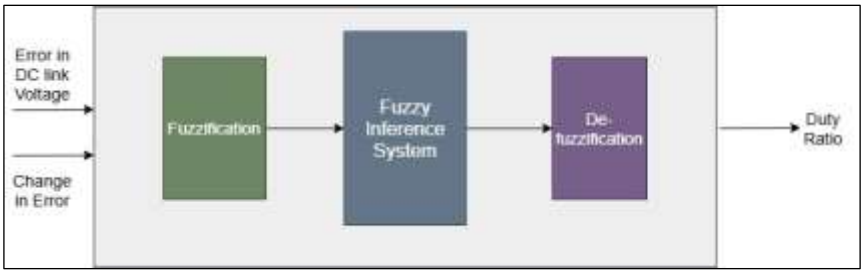


Figure 2: Working of Fuzzy logic controller for HEV.
Source: Authors, (2026).

Figure 2 illustrates the working principle of the FLC. The controller receives two input variables: the error in the DC link voltage and the change in this error. The process begins with fuzzification, where the fuzzifier converts these crisp input values into corresponding fuzzy sets. Next, the fuzzy inference system (FIS) applies a set of predefined fuzzy rules to evaluate the fuzzified inputs and generate an intermediate fuzzy output. Finally, this output is passed through the defuzzification stage, which converts it back into a crisp value representing the duty ratio, serving as the final control signal for the system.

III.1 MEMBERSHIP FUNCTIONS

The input is classified into 7 types that are negative big (NB), negative medium (NM), negative small (NS), zero (ZE), positive small (PS), positive medium (PM) and positive big (PB) [19]. Error in dc link voltage value in range between -200 to +200. Similarly, the change in error input function also 7 membership functions for each input. The range is taken -1 to +1 and the duty ratio output function also 7 membership function and range taken between -100 to +100. Table 1 presents the rule-based approach for managing various input value ranges in the fuzzy charge control system.

Table 1: Rule strategy for Duty ratio.

Duty Ratio		**Error**						
		NB	**NM**	**NS**	**Z**	**PS**	**PM**	**PB**
Change in Error	**PS**	Z	PS	PM	PB	PB	PB	PB
	PM	NS	Z	PS	PM	PB	PB	PB
	PB	NM	NS	Z	PS	PM	PB	PB
	Z	NB	NM	NS	Z	PS	PM	PB
	NS	NB	NB	NM	NS	Z	PS	PM
	NM	NB	NB	NB	NM	NS	Z	PS
	NB	NB	NB	NB	NB	NM	NS	Z

Source: Authors, (2026).

IV. ANN CONTROLLER FOR HYBRID SYSTEM

ANNs have become increasingly popular and are now among the most commonly used learning algorithms for identifying complex relationships within datasets. In this paper, ANN is utilized to improve the estimation of the battery's SOC [20]. Due to their strong processing capabilities, ANNs can deliver highly satisfactory results, particularly in complex pattern recognition problems. The principle of an ANN is inspired by the structure and function of the human brain. In a biological brain, a large number of neurons are interconnected, forming a network that processes inputs received from sensory receptors. Similarly, in ANN, artificial neurons are interconnected, with each neuron belonging to a specific layer within the network [21]. From figure 3 the general architecture of the ANN used in this work consists of three layers input layer,hidden layer and the output layer. Each neuron in one layer is connected to all neurons in the subsequent layer through weighted connections. The process begins when neurons in the input layer receive input values and pass them to the neurons in the hidden layer. The hidden layer processes these values, and its number of layers defines the depth of the network. Finally, the output layer generates the final output of the network. During this process, each neuron applies a mathematical operation that combines the weighted sum of its inputs through an activation function, enabling the ANN to perform nonlinear transformations and learn complex mappings between inputs and output [22], [23].

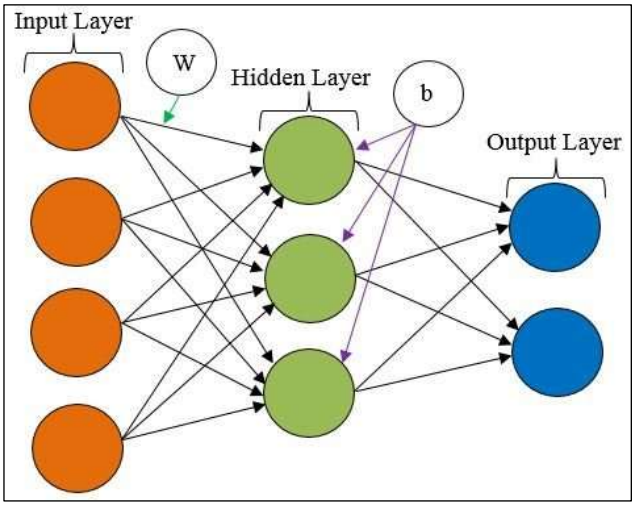


Figure 3: Architecture of ANN.
Source: Authors, (2026).

In this application, ANN is an effective method for improving SOC estimation, which requires modeling a nonlinear function. The hidden layers in an ANN learn to map the relationships between inputs and outputs, enabling the network to represent both linear and nonlinear functions.

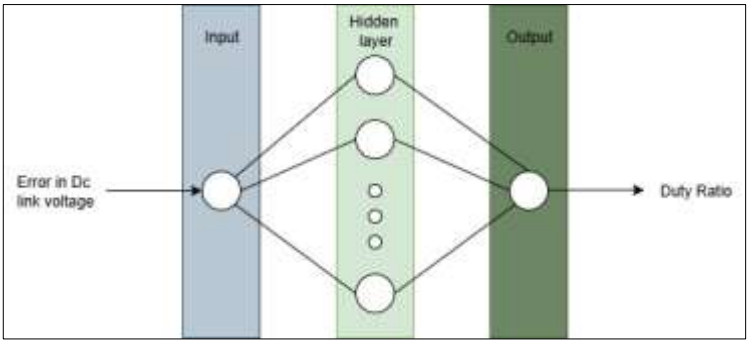


Figure 4: ANN based SOC controller.
Source: Authors, (2026).

V. RESULTS AND DISCUSSIONS

The key advantage of such systems is their adherence to precise rules, ensuring execution within the specified rule limits. Utilizing the dc link voltage battery SOC can enhance, enabling the hybrid mode for longer-distance travel. The Simulink model used to simulate PID controller, FLC and ANN controller is represented in Figure 5 and Figure 6.

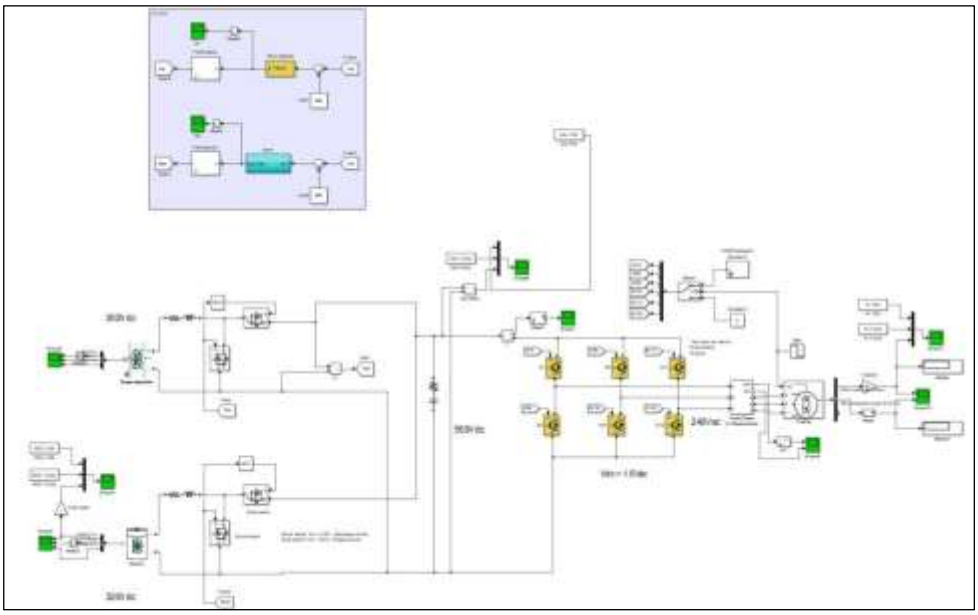


Figure 5: Simulink Model of Hybrid System.
Source: Authors, (2026).

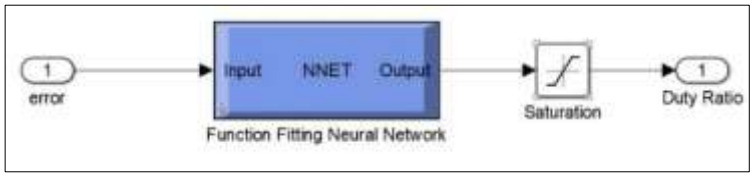


Figure 6: ANN controller of Hybrid System.
Source: Authors, (2026).

The preceding section's model is simulated and the results are provided in the following two conditions, that are:

- At starting condition and
- At braking condition

At starting condition both battery and UC discharges and three phase converter works as a rectifier which is connected to the PMSM drive. From 0 sec to 4 sec considered as the starting and normal driving of the vehicle. Braking applied at 4 sec then motor works like a generator and three phase converter works as an inverter then battery stores energy by using regenerative braking method. From Figure 7 the SOC curve of battery with PID controller is initially at 100% and gradually decreases to 99.892% up to braking condition of 4 sec. At the time of braking condition, the motor will work as a generator and charges the battery from 4 sec to 5 sec from 99.892% to 99.896%. With FLC the SOC curve of the battery is initially at 100% and gradually decreases to 99.946% up to braking condition of 4 sec. At the time of braking condition, from 4 sec to 5 sec battery charges from 99.946% to 99.952%. With ANN controller the SOC curve of battery is initially at 100% and gradually decreases to 99.922% up to braking condition of 4 sec. At the time of braking condition, from 4 sec to 5 sec battery charges from 99.922% to 99.928%. we can conclude that by using FLC battery SOC maintained smoothly and driving range also increases at the end.

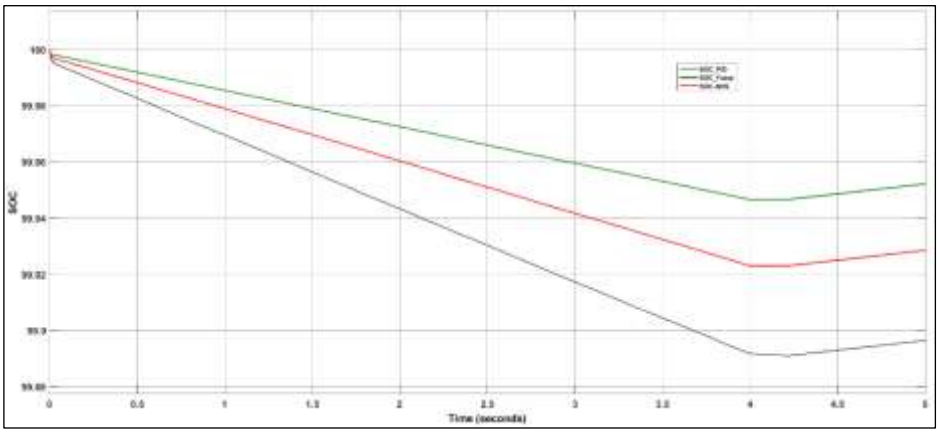


Figure 7: Battery SOC using PID and Fuzzy controller.
Source: Authors, (2026).

Figure 8 shows dc link voltage of PID controller, fuzzy and ANN controller. Using an ANN controller reduces ripples more effectively than a fuzzy controller. By minimizing ripples in the DC-link voltage, the voltage error is decreased and the SOC is improved.

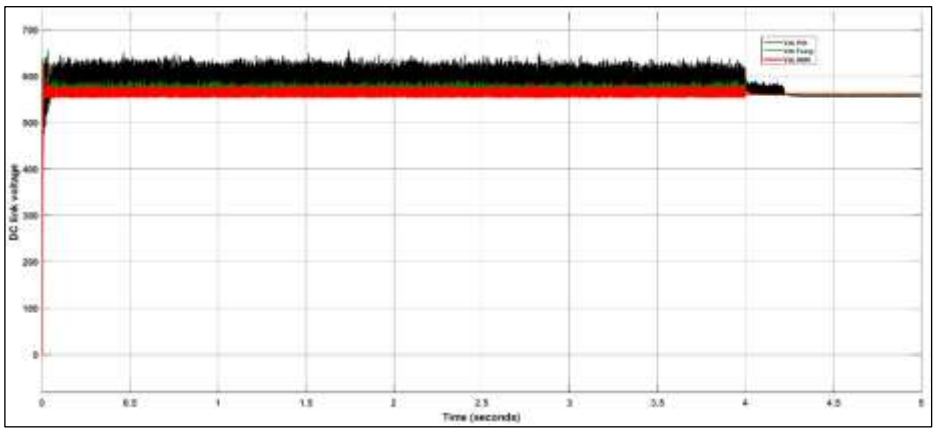


Figure 8: DC link voltage of PID, Fuzzy and ANN controller.
Source: Authors, (2026).

Figure 9 shows the speed characteristics of PID, FLC and ANN controller, where the simulation implemented with same speed and torque parameters but there is a smooth operation with Fuzzy and ANN controller.

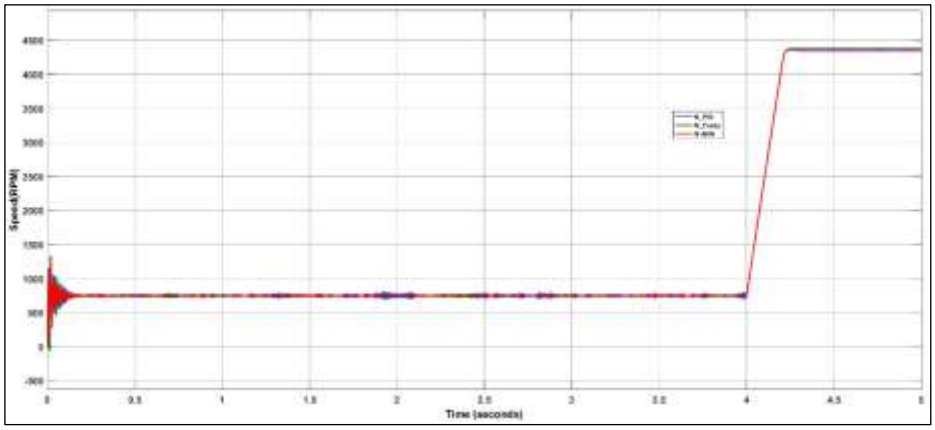
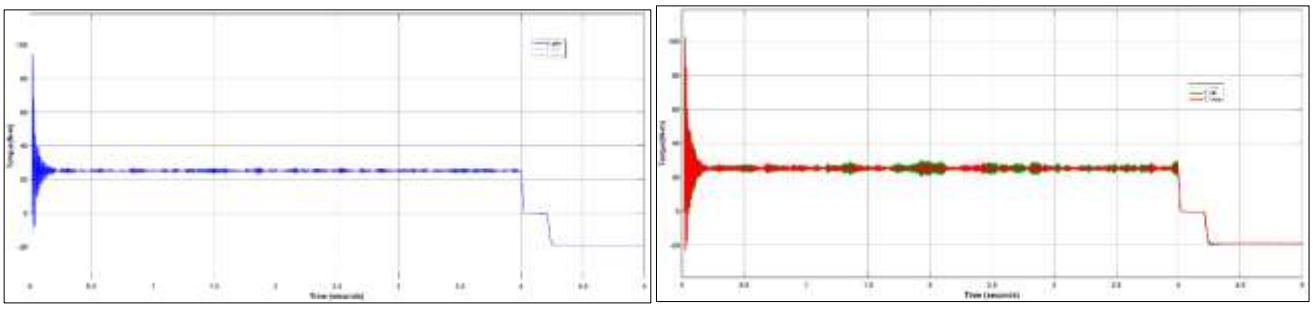


Figure 9: Speed characteristics of PID, Fuzzy and ANN controller.
Source: Authors, (2026).

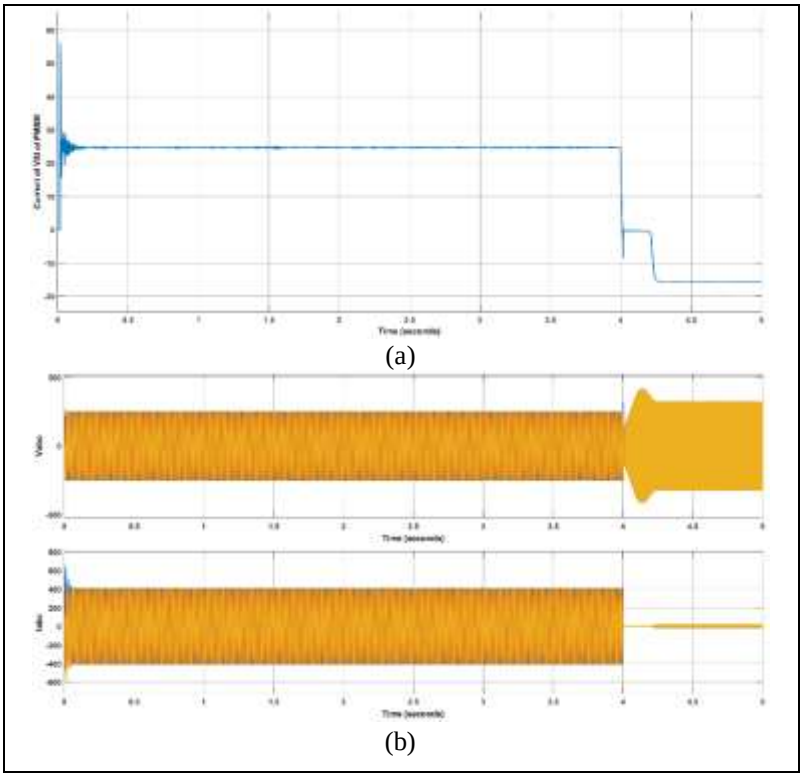
Figure 10 shows the torque characteristics of PID, FLC and ANN controller, where the simulation implemented with same speed and torque parameters but ripples reduced with FLC and ANN controller. At the braking condition torque to the negative condition and battery charges using inverter.



(a) (b)

Figure 10: (a) Torque characteristics of ANN controller and (b) Torque characteristics of PID and Fuzzy controller.
Source: Authors, (2026).

Figure 11 shows the DC link current and Vabc and Iabc of VSI of PMSM. At braking condition current goes to the negative peak.



(a)

(b)

Figure 11: (a) DC link Current of PMSM and (b) V_{abc} and I_{abc} of VSI of PMSM.
Source: Authors, (2026).

Figure 12 indicates Battery and UC SOC, voltage and current waveforms with ANN controller.

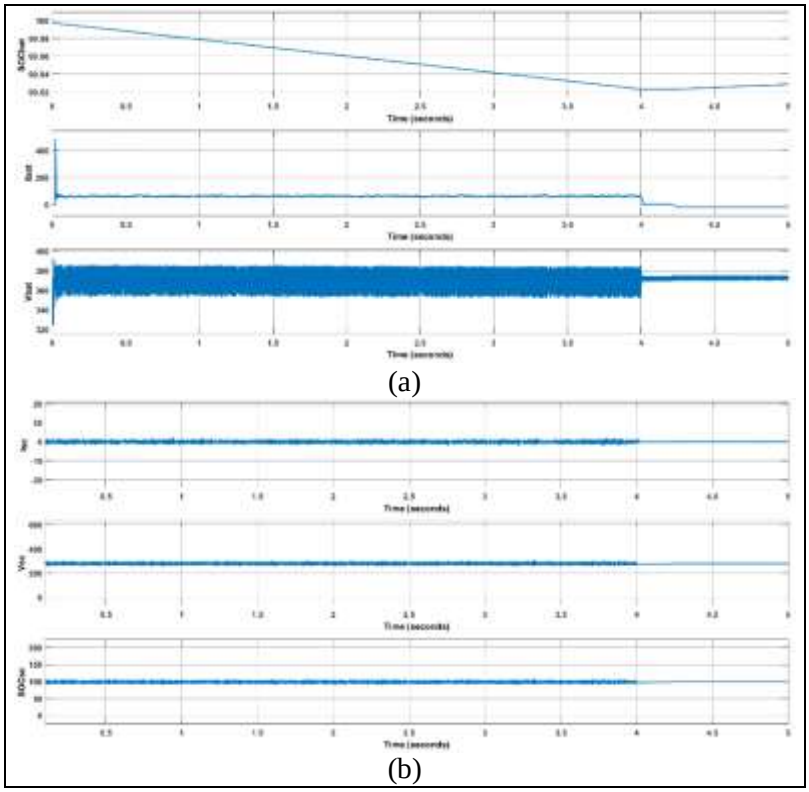


Figure 12: SOC, current and voltage with ANN controller of (a) Battery (b) Ultracapacitor.
Source: Authors, (2026).

Table 2 below presents a comparison of various parameters between the PID, FLC and ANN controller.

Table 2: Comparison of various parameters using different Energy management systems.

Parameter	Battery Controller		
	PID	ANN	FLC
Time(sec)	5	5	5
Initial SOC%	100	100	100
Final SOC%	99.896	99.928	99.952
SOC drop	Sharp	Small	Small

Source: Authors, (2026).

The findings indicate that the FLC and ANN controller are more advanced and precise systems, grounded on the input/output data pairs of the system. The objectives of the study were fully met, with simulation results confirming that the fuzzy controller can improve SOC more effectively than the ANN controller. These outcomes not only validate the potential of FLC for real-world battery control applications but also provide a more comprehensive benchmark for future controller evaluations.

VI. CONCLUSIONS

The PID controller regulates the battery SOC, optimizing it for extended driving ranges. Meanwhile, the Fuzzy Logic Controller utilizing input output data pairs, further enhances SOC management. To enable a comprehensive comparison, advanced optimization techniques including the PID, FLC and ANN controllers were implemented and evaluated. Simulation results reveal that the PID controller maintained the battery SOC at 99.89% by the end of the drive cycle, though minor drops occurred during acceleration phases. The ANN controller achieved a slightly higher SOC of 99.92%, while the fuzzy controller, employing a triangular membership function, sustained an even higher SOC above 99.952% at the end of the drive cycle. This outcome highlights the superior capability of the fuzzy controller in preserving SOC stability, maintaining approximately 99.952% even after a 5-second drive.

It can be concluded that the implementation of ANN and Fuzzy controllers enhances battery performance, improving the SOC stability and enabling HEV to achieve longer driving ranges. Overall, both the ANN and FLC approaches significantly enhanced battery performance, improving SOC stability and thereby extending the vehicle's operational range. The study demonstrates how integrating machine learning techniques with traditional power electronic control models can effectively improve battery management and overall system efficiency an advancement particularly relevant for EV applications. As HEV continue to shape the future of transportation, this work underscores the importance of intelligent, energy-efficient control systems that support the shift toward sustainable mobility.

VII. REFERENCES

[1] S.L. Sreedevi and B.T. Geetha, “Optimal EV Charging Using ANT Lion Optimized ANN Based MPPT with High Gain Modified Zeta Converter,” SSRG International Journal of Electrical and Electronics Engineering, vol. 11, no. 2, pp. 24-36, February 2024. doi: 10.14445/23488379/IJEEE-V11I2P104

[2] Fengqi Zhang, Lihua Wang, Serdar Coskun, Hui Pang, Yahui Cui and Junqiang Xi, “Energy Management Strategies for Hybrid Electric Vehicles: Review, Classification, Comparison, and Outlook,” Energies, vol. 13, no. 3352, pp. 1-36, 2020. doi: 10.3390/en13133352

[3] Ivana Jovanovic, Fardis Nakhaei, Daniel Kržanovic, Vesna Conic and Daniela Uroševic, “Comparison of Fuzzy and Neural Network Computing Techniques for Performance Prediction of an Industrial Copper Flotation Circuit,” Minerals, vol. 12, no. 1493, pp. 1-50, November 2022. doi: 10.3390/min12121493

[4] Kaif Ahmed Lodi, Abdul R. Beig, Khaled Ali and Zeyar Aung, “ANN-Based Improved Direct Torque Control of Open-End Winding Induction Motor,” IEEE Transaction on Industrial Electronics, vol. 71, No. 10, pp. 12030-12040, October 2024. doi: 10.1109/TIE.2024.3357865

[5] Thomas Guillod, Panteleimon Papamanolis and Johann W. Kolar, “Artificial Neural Network (ANN) Based Fast and Accurate Inductor Modeling and Design,” IEEE Transaction, vol. 1, pp. 284-299, July 2020. doi: 10.1109/OJPEL.2020.3012777

[6] Zhumu Fu, Haocong Wang, Fazhan Tao, Baofeng Ji, Yongsheng Dong and Shuzhong Song “Energy Management Strategy for Fuel Cell/Battery/Ultracapacitor Hybrid Electric Vehicles Using Deep Reinforcement Learning With Action Trimming,” IEEE Transaction, vol. 71, no. 7, pp. 7171-7185, July 2022. doi: 10.1109/TVT.2022.3168870

[7] Amit kumer podder and Oishikha chakraborty “Control Strategies of Different Hybrid Energy Storage Systems for Electric Vehicles Applications,” IEEE Transaction, vol. 9, pp. 51865-51895, 2021. doi: 10.1109/ACCESS.2021.3069593

[8] Chokri Ben Salah and Mohamed Ouali, "Comparison of fuzzy logic and neural network in maximum power point tracker for PV systems," Electric Power Systems Research, vol. 81, no.1, pp.43-50, January, 2021. doi: 10.1016/j.epsr.2010.07.005

[9] Naoui Mohamed, Flah Aymen, Abdullah Altamimi, Zafar A. Khan and Sbita Lassaad, "Power Management and Control of a Hybrid Electric Vehicle Based on Photovoltaic, Fuel Cells, and Battery Energy Sources," Journal of Sustainability, vol. 14, no.2551, pp.1-20,2022. doi: 10.3390/ su14052551

[10] Trinh Cong Truong, Thanh Nguyen Vu, Nam Le Hai, Thang Nguyen Duy, Duong Doan Le Quy, Vuong Dang Quoc, Duc-Quang Nguyen and Bao Doan Thanh, “Optimal Electromagnetic Parameters of SPMSM for Electric Vehicles Based on Genetic Algorithm Technique,” International Journal on Electrical Engineering and Informatics, vol. 16, no. 1, pp. 137-148, March,2024. doi: 10.15676/ijeei.2024.16.1.9

[11] Pengli Yu, Mince Li and Yujie Wang and Zonghai Chen “Fuel Cell Hybrid Electric Vehicles: A Review of Topologies and Energy Management Strategies,” World Electric Vehicle Journal, vol. 13, no. 172, pp. 1-19, 2022. doi: 10.3390/wevj13090172

[12] Mohammad suhail, iram akhtar, sheeraz kirmani and mohammed jameel, “Development of Progressive Fuzzy Logic and ANFIS Control for Energy Management of Plug-In Hybrid Electric Vehicle,” IEEE Transaction, vol 9, pp. 62219-62231, 2021. doi: 10.1109/ACCESS.2021.3073862

[13] R. Ranjith Kumar, C. Bharatiraja and K. Udhayakumar,” Advances in Batteries, Battery Modelling, Battery Management System, Battery Thermal Management, SOC, SOH, and Charge/ Discharge Characteristics in EV Applications,” IEEE Transaction, vol. 11, pp. 105761-105809, 2023. doi: 10.1109/ACCESS.2023.3318121

[14] Chandra Sekhar Goli, Somasundaram Essakiappan and Prasanth Sahu, “Review of Recent Trends in Design of Traction Inverters for Electric Vehicle Applications,” IEEE 12th International Symposium on Power Electronics for Distributed Generation Systems (PEDG), Chicago, USA, pp. 1-6, 2021. doi: 10.1109/PEDG51384.2021.9494164

[15] Kiran Kumar and G. Tulasi Ram Das, ” Adaptive fuzzy controller based self regulated reference stator flux estimator of direct torque control for three level inverter fed IPMSM,” International Journal of Intelligent Engineering and Systems, vol. 13, no. 2, pp. 11–19, 2020. doi: 10.22266/ijies2020.0430.02

[16] Hadjer Abderrezek and Ameur Aissa, “Extended Kalman Filter Estimator and Mamdani Fuzzy Logic Controller for Sensorless DC Motor Speed Control,” Journal of Engineering and Technology for Industrial Applications, vol. 11, no.55, pp. 11-19, October, 2025. doi: 10.5935/jetia.v11i55.1836

[17] Walid S. E. Abdellatif, Mohamed Saad Mohamed, Shimaa Barakat and Ayman Brisha, “A Fuzzy Logic Controller Based MPPT Technique for Photovoltaic Generation System,” International Journal on Electrical Engineering and Informatics, vol. 13, no.2, pp. 394-417, June, 2021. doi: 10.15676/ijeei.2020.13.2.9

[18] S Devikala, Rabi. J, V.P. Murugan, J. S. Christy Mano Raj, K. Mohana Sundaram and K. Sivakumar,” Development of Fuzzy Logic Controller in Automatic Vehicle Navigation using IoT,” Journal of Electrical Systems, vol. 20, no. 3, pp. 114–121, 2024. doi: 10.52783/jes.1254

[19] Hari maghfiroh, Oyas Wahyunggoro and Adha Imam Cahyad, “Energy Management in Hybrid Electric and Hybrid Energy Storage System Vehicles: A Fuzzy Logic Controller Review,” IEEE Transaction, vol. 12, pp. 56097-56109, 2024. doi: 10.1109/ACCESS.2024.3390436

[20] Mehmet Kurucan, Mete Ozbaltan, Zeki Yetgin and Alkan Alkaya, “Applications of artificial neural network based battery management systems: A literature review,” Renewable and Sustainable Energy Reviews, vol. 192, pp. 1-13, March, 2024. doi: 10.1016/j.rser.2023.114262

[21] Soumya Sathyan, V Ravikumar Pandi, Ashin Antony, Surender Reddy Salkuti and Preetha Sreekumar, “ANN-based energy management system for PV-powered EV charging station with battery backup and vehicle to grid support,” International Journal on Green Energy, vol. 21, no.6, pp. 1279-1294, August, 2023. doi: 10.1080/15435075.2023.2246048

[22] Bikash Sah; Praveen Kumar and Sanjay Kumar Bose,” A Fuzzy Logic and Artificial Neural Network-Based Intelligent Controller for a Vehicle-to-Grid System,” IEEE Transaction, vol. 15, no. 3, pp. 3301-3311, July, 2020. doi: 10.1109/JSYST.2020.3006338

[23] N. Senthil Kumar, V. Sadasivam and H. M. Asan Sukriya, “A Comparative Study of PI, Fuzzy, and ANN Controllers for Chopper-fed DC Drive with Embedded Systems Approach,” Electric Power Components and Systems , vol. 36, no. 7, pp.680-695, 2008. doi: 10.1080/15325000701881944

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 348-355. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2983

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

EVALUATING THE MEAN TIME BETWEEN FAILURE (MTBF) OF CRITICAL BUILDING SYSTEMS IN STUDENTS' HOSTEL BUILDINGS WITHIN LAGOS METROPOLIS

Olajide Julius Faremi*[1], Damilola Abdurazaq Lukman[2] and Dele Roger Simeon[3]

[1, 2, 3] Department of Building, Faculty of Environmental Sciences, University of Lagos, Lagos State, Nigeria.

[1]https://orcid.org/0000-0002-1097-2087, [2]https://orcid.org/0009-0000-9288-0161, [3]https://orcid.org/0000-0002-3927-7547

Email: *ofaremi@unilag.edu.ng, damilolaoshorinde@gmail.com, dsimeon@unilag.edu.ng

ARTICLE INFO

Article History
Received: November 20, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Building systems,
Lagos Metropolis,
Maintenance management,
Mean Time Between Failure,
Students Hostels.

ABSTRACT

Student hostel buildings in Lagos Metropolis, Nigeria, serve as essential infrastructure for higher education, yet they face significant maintenance challenges that impact functionality, safety, and student comfort. This study evaluates the Mean Time Between Failures (MTBF) for critical building systems in students' hostels across four types of institutions. Using purposive sampling approach, data were collected from 56 maintenance staff and porters. The data were analysed using mean, relative influence index, Mann-Whitney U test and Kruskal-Wallis test as descriptive and inferential statistical tools, respectively. The results show that MTBF is highest for foundation issues (MTBF = 17182.7 hours) and lowest for wiring systems (MTBF = 1304.9 hours), while budget constraints (RII = 0.79), inadequate monitoring systems (RII = 0.77), spare parts availability (RII = 0.76), and poor maintenance documentation (RII = 0.76) are the topmost factors influencing MTBF in hostel buildings. The study concludes that structural and foundational elements are the most reliable systems in students' hostels with rare occurrences of failures; while wiring systems, lighting bulbs and water pipes are vulnerable systems with relatively low reliability. The study recommends improved budget allocation and technology integration as strategic interventions for enhancing system reliability and students' satisfaction in hostel buildings.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Student accommodation in higher institutions plays a pivotal role in supporting academic performance, health, and overall well-being. In Lagos Metropolis, Nigeria's economic and educational hub, the demand for quality hostel facilities has surged with increasing student populations. Yet, maintenance management remains inadequate, leading to frequent system failures and reduced functionality [1], [2]. Maintenance performance indicators, such as Mean Time Between Failures (MTBF), are essential for quantifying system reliability and guiding resource allocation in educational infrastructure [3], [4]. MTBF measures the average time a system operates before failure, providing insights into durability and maintenance needs. In the context of student hostels, critical building systems - such as electrical, plumbing, structural, fire safety, waste disposal, sewage, solid waste management, CCTV, internet/Wi-Fi, and lightning protection - must function reliably to ensure safety and comfort [5], [6]. However, studies highlight persistent issues like reactive maintenance, budget constraints, and a lack of standardized metrics in Nigerian hostels, resulting in extended downtimes and health risks [7], [8]. This study addresses a key gap by measuring average MTBF values for critical building systems across Lagos hostels. The objectives of the study are:

- To evaluate the mean time between failure of critical systems in students' hostel ;
- To assess factors influencing mean time between failure of critical systems in students' hostel.

The research hypotheses formulated for the study are:

- Ho1: There is no statistical difference in the perception of maintenance staff and porters on the factors influencing failure of critical building systems in students' hostels.
- Ho2: There is no statistical difference in the factors influencing failure of critical building systems in students' hostels across the various types of institutions.

By focusing on MTBF, the study contributes to evidence-based guidelines for improving maintenance in educational facilities, aligning with global calls for sustainable infrastructure [9], [10].

II. THEORETICAL REFERENCE

II.1 MAINTENANCE PERFORMANCE INDICATORS IN BUILDINGS

Maintenance performance indicators (KPIs) are metrics used to evaluate the effectiveness and efficiency of maintenance operations in buildings [3]. These indicators help facility managers assess system reliability, optimize resource use, and reduce costs. Key KPIs include Mean Time Between Failures (MTBF), Mean Time To Repair (MTTR), and Overall Equipment Effectiveness (OEE) [11]. In educational facilities like student hostels, these metrics are crucial for ensuring the continuous operation of critical systems, which directly impacts student well-being and academic performance [12], [5]. Literature emphasizes the role of KPIs in transitioning from reactive to preventive maintenance. Reactive approaches, common in Nigerian hostels, lead to high costs and downtimes [13], [14]. Preventive strategies, guided by KPIs, can extend system life and improve safety [15]. For instance, MTBF helps predict failure patterns, allowing timely interventions [16].

II.2 MTBF IN HOSTEL BUILDINGS

Mean Time Between Failures (MTBF) is defined as the average operational time between failures, calculated as total uptime divided by the number of failures [17]. It serves as a reliability metric, with higher values indicating more durable systems. In building maintenance, MTBF is applied to systems like electrical wiring, plumbing, and structural elements to assess performance [18], [19]. In student hostels, MTBF is particularly relevant due to high usage and limited budgets. Studies show low MTBF for electrical systems due to overloads and poor wiring, leading to frequent outages [20]. Structural systems, like foundations, typically have higher MTBF as they are less prone to daily wear [21]. In Nigerian contexts, MTBF is often low in public institutions due to inadequate monitoring and funding [9], [22]. Research gaps include limited empirical data on MTBF in Lagos hostels and comparisons across institution types [23]. Table 1 summarizes recent literature on MTBF in building maintenance.

Table 1: Summary of recent literature on MTBF in building maintenance.

Authors	Focus	Methods	Limitations
[17]	Explanation and standards for MTBF in equipment reliability.	Review	Focused on general building equipment and not specific to students' hostels.
[4]	Analysis of downtime and reliability in hostel building maintenance.	Case study	Limited to one institution and did not provide for institutional comparisons.
[18]	Quality control analysis of downtime and MTBF for water supply pipes.	Analysis	The study was restricted to plumbing systems.
[19]	Implementation of MTBF in logistical support for operations.	Case study	The study was not specific to education instituional buildings.
[3]	KPIs including MTBF for maintenance operations.	Review	The study provided broad Key Performance Indicators for maintenance and was not focused on student hostel buildings.

Source: Authors (2026).

From Table 1, existing literature on MTBF applications are limited in scope, with few studies on educational facilities. This study addresses this by measuring MTBF across critical systems in various types of tertiary institutional students' hostels.

II.3 CHALLENGES IN MEASURING MTBF IN EDUCATIONAL FACILITIES

Measuring MTBF in a hostel faces challenges like data inaccuracy, lack of reporting systems, and resource constraints [24]. In Nigeria, reactive maintenance hinders accurate failure tracking [7]. Institutional disparities, with private universities having better data systems, affect MTBF calculations [25].

III. MATERIALS AND METHODS

This study was conducted in Lagos, targeting student hostels in four institution types: Federal University, Federal Polytechnic, Private University, and Federal College of Education. A mixed-methods approach was used, with quantitative data from questionnaires focused on MTBF. The population comprised maintenance staff and porters responsible for the daily operations and maintenance of systems and services in students' hostel buildings. Purposive sampling selected 56 respondents with direct experience. Questionnaire was the instrument used in collecting primary data for the study with a section of the questionnaire eliciting information on the number of failures for critical systems (e.g., wiring repairs, foundation issues) in the last one year. MTBF was calculated as:

MTBF = Total Operational Time / Number of Failures
Assuming 24/7 operation.

Data were analyzed using descriptive statistics (means, frequencies) in SPSS. Comparisons across institutions used Mann-Whitney U test to test the hypothesis of no significant difference in perceptions. Validity was ensured through pilot testing and expert review; reliability via Cronbach's Alpha (>0.7).

IV. RESULTS AND DISCUSSIONS

This section presents the findings of the study, organized into four subsections: respondent demographics, MTBF for critical building systems, factors influencing failures of building systems, statistical comparisons between institutions, and a thematic discussion of the results. The analysis draws from quantitative data collected through structured questionnaires distributed to maintenance staff and porters in maintenance management at students' hostels. Both descriptive and inferential statistical techniques were employed to identify key constraints affecting MTBF of critical systems and to determine whether significant differences exist between institutional contexts.

IV.1 RESPONDENTS' PROFILE

Understanding the background of the respondents provides critical context for interpreting the results. Table 2 summarizes the 56 respondents: 71.4% male, 28.6% female; roles: 53.6% maintenance staff, 46.4% porters; experience: over 10 years (32.1%), 7-10 years (25%). Institutions: Federal University (35.7%), Federal Polytechnic (28.6%), Private University (21.4%), Federal College of Education (14.3%).

Table 2: Demographic profile of respondents.

Profile	N	Percent	Cumulative Percent
Gender			
Male	31	55.4	55.4
Female	25	44.6	100
Role			
Maintenance Staff	27	48.2	48.2
Porter	29	51.8	100
Age Group			
16-24 years	0	0	0
25-34 years	5	8.9	8.9
35-44 years	24	42.9	51.8
45-54 years	23	41.1	92.9
55+ years	4	7.1	100
Experience			
Less than 1 year	0	0	0
1-3 years	13	23.2	23.2
4-6 years	30	53.6	76.8
7-10 years	12	21.4	98.2
Over 10 years	1	1.8	100
Institution Type			
Federal University	21	37.5	37.5
Federal Polytechnic	10	17.9	55.4
Private University	15	26.8	82.2
Federal College of Education	10	17.9	100
Duration of Working			
Less than 6 months	0	0	0
6-12 months	2	3.6	3.6
1-2 years	18	32.1	35.7
3-5 years	18	32.1	67.8
Over 5 years	18	32.1	100

Source: Authors (2026).

The gender distribution shows that 55.4% of the respondents were male, while 44.6% were female. This shows that the distribution gives an adequate representation of male and female respondents. Regarding role type, the table shows that about 51.8% of the respondents were porters, while about 48.2% were maintenance staff. This shows that the distribution was evenly shared amongst the respondents. In terms of age category, 42.9% of maintenance staff/porter respondents fall within the age range of 35-44 years, and 41.1% of the respondents fall within the range of 45-54 years. Only a few numbers of respondents, representing 8.9% fall within the range 25-34 years. 53.6% of the respondents have a working experience of 4-6 years in hostel maintenance/management, and about 23.2% and 21.4% of the respondents have a working experience of 1-3 years and 7-10 years, respectively, in hostel maintenance. Regarding institutional affiliation, 37.5% of the respondents were from Federal University, 26.8% of the respondents from Private University, and 17.9% of the respondents from both Federal Polytechnic and Federal College of Education. In terms of how long the respondents have spent working in the hostel, 32.1% of the respondents have spent between 1 and 2 years, 3-5 years, and over 5 years in the hostel, respectively, which shows an even distribution of experience by the respondents. Only a few numbers of respondents representing 3.6% have worked in the hostel for less than 6 - 12 months.

IV.2 MTBF VALUES FOR CRITICAL BUILDING SYSTEMS IN STUDENTS' HOSTEL BUILDINGS

Following the demographic overview, this subsection presents the MTBF values for critical building systems in students' hostel buildings. Drawing responses from 4 institutions, the data highlight the MTBF values of critical systems in students' hostel. Table 3 presents a ranked list of MTBF values of various critical systems in students' hostels.

Table 3: MTBF values for critical building systems in students' hostel buildings.

Critical systems in hostels	FU (N=21)		FP (N=10)		PU (N=15)		FCE (N=10)		Total (N=56)	
	MTBF	Rank	MTBF	Rank	MTBF	Rank	MTBF	Rank	MTBF	Rank
Wiring systems	1303.7	1	1619.2	3	1247.0	2	1080.1	1	1304.9	1
Lighting systems (bulbs)	1335.0	2	1598.7	2	1228.9	1	1228.5	3	1334.7	2
Water pipes	1457.4	3	1461.0	1	1338.1	3	1126.3	2	1366.9	3
Sockets and switches	1659.6	4	1648.8	4	1546.2	5	1322.3	4	1567.3	4
Shower/bathing facility	1732.1	5	1739.5	5	1699.2	7	1435.0	5	1671.6	5
Drainage	1836.4	6	2555.0	8	1485.7	4	1581.0	7	1825.2	6
Toilet/bathroom fixtures	1937.4	7	2000.2	6	1717.9	8	1649.8	8	1838.5	7
Power distribution panels	2044.0	8	2803.2	9	1613.0	6	1437.1	6	1955.7	8
Door/window frames	2217.8	9	2175.4	7	2423.6	9	1839.6	9	2197.8	9
Backup power systems	2753.1	10	3124.4	10	2550.1	10	1912.6	10	2615.0	10
Walls	3476.2	11	3577.0	11	3212.0	11	2482.0	11	3245.9	11
Floor finishes	3476.2	12	3825.2	12	3260.7	12	2847.0	12	3368.4	12
Staircase and railing	4588.6	13	4701.2	13	3990.7	13	3942.0	13	4333.1	13
Roofs	5318.6	14	6307.2	14	5256.0	14	5256.0	14	5467.2	14
Water heaters	7230.5	15	8059.2	16	5548.0	15	7884.0	16	7044.5	15
Building foundation	17647.6	16	17519.0	15	16132.0	16	17446.0	15	17182.7	16

NOTE: FU is Federal University, FP is Federal Polytechnic, PU is Private University, FCE is Federal College of Education.
Source: Authors (2026).

The result shows that the Mean Time Between Failure (MTBF) values for critical building systems in students' hostels across the four higher institutions in Lagos Metropolis varies in reliability. Wiring systems ranks 1st with the lowest overall MTBF of 1304.9 hours and, indicating that it is the least reliable in the systems category with the most frequent failures. The frequent failure of wiring systems could potentially be due to electrical overloads from multiple devices, substandard wiring materials, or unstable power supply, which could lead to safety hazards like fire outbreak or supply outages affecting students' comfort and productivity. Lighting systems (bulbs) ranks 2nd with a MTBF of 1334.7 hours, indicating frequent bulb failures. This could be as a result of voltage fluctuations, poor quality lamps, or excessive demand in communal areas, underscoring the need for energy-efficient upgrades to ensure consistent lighting for study and navigation in the hostels. Water pipes rank 3rd with a MTBF of 1366.9 hours, highlighting recurrent plumbing leaks as a major issue and plumbing pipes as a key vulnerable system for failure, sockets/switches rank 4th with a MTBF of 1567.3 hours, suggesting moderate reliability for electrical sockets and switches, with failures often linked to wear and tear from frequent use, misuse and poor handling by the students.

The result also shows that systems with lower failure rates include; water heaters, which rank 15th with a MTBF of 7044.5 hours, demonstrating high reliability, possibly due to infrequent use or the unavailability of the system in most hostel buildings. Foundation issues rank 16th with the highest overall MTBF of 17182.7 hours, indicating that the foundations of hostel buildings rarely exhibit any form of failure due to their engineered stability as any issue could have severe implications in terms of building stability and safety. Comparatively, the result shows that in the Federal University, wiring systems has the lowest MTBF of 1303.7 hours, revealing that electrical repairs is a primary concern in the students' hostels. Other systems in the Federal University students' hostels having low reliability include; lighting systems (MTBF = 1353.0 hours), water pipe leaks (MTBF = 1457.4 hours), Sockets and switches (MTBF = 1659.6 hours), Shower/bathing facility (MTBF = 1732.1 hours), In the Federal Polytechnic however, systems with low reliability include; water pipe leaks (MTBF = 1461.0 hours), lighting systems (MTBF = 1598.7 hours), wiring systems (MTBF = 1619.2 hours), and socketets and switches (MTBF = 1649.8 hours).

Furthermore, the result shows that systems with low reliability in Private University are; lighting system (MTBF = 1228.9 hours), wiring systems (MTBF = 1247.0 hours), water pipes (MTBF = 1247.0) and drainage (MTBF = 1485.7 hours) while the systems with low reliability in Federal College of Education are; wiring systems (MTBF = 1080.1 hours), water pipes (MTBF = 1126.3 hours), lighting systems (MTBF = 1228.5 hours) and sockets and switches (MTBF = 1323.2 hours). In general, public institutions such as Federal University and Federal Polytechnic exhibit higher MTBF in structural components, suggesting better long-term durability, while Private Universities and Federal Colleges of Education show more variability in systems reliability, with private institution underperforming in major structures and colleges in minor fixtures. The results highlight potential disparities in maintenance investment and usage patterns across institutional types. The findings show that the variability from low MTBF in electrical and plumbing related systems to high MTBF in structural systems align with usage intensity, suggesting a need for improved resource allocation for maintenance management activities aimed at enhancing reliability, reduced operational costs, and improve students' comfort and satisfaction in the hostels.

IV.3 FACTORS INFLUENCING FAILURE OF BUILDING SYSTEMS AND COMPONENTS IN STUDENTS' HOSTELS

The respondents were also asked to rate the extent to which they agree that the under-listed factors influence failures of building systems and components in students' hostels. Table 4 presents a ranked list of 15 identified factors, analyzed using the Relative Influence Index (RII) to assess the severity of each factor from the perspective of maintenance staff and porters across the four institutions.

Table 4: Factors influencing failures of critical building systems in students' hostels.

Factors influencing critical building systems failure	Maintenance Staff (N=27)			Porter (N=29)			Total (N=56)		
	RII	Rank	Remark	RII	Rank	Remark	RII	Rank	Remark
Budget constraints	0.76	4	EI	0.83	1	EI	0.79	1	EI
Inadequate monitoring systems	0.76	4	EI	0.78	2	EI	0.77	2	EI
Spare parts availability	0.75	7	I	0.77	3	EI	0.76	3	EI
Poor maintenance documentation	0.75	7	I	0.77	3	EI	0.76	3	EI
Poor installation quality	0.78	1	EI	0.73	6	I	0.75	5	I
Inadequate preventive maintenance	0.77	3	EI	0.72	8	I	0.75	5	I
Excessive user load and misuse	0.76	4	EI	0.7	10	I	0.73	7	I
Poor communication channels	0.78	1	EI	0.68	13	I	0.73	7	I
Lack of user education	0.73	9	I	0.72	8	I	0.73	7	I
Design inadequacy	0.70	11	I	0.77	3	EI	0.73	7	I
Building age and degradation	0.70	11	I	0.73	6	I	0.72	11	I
Climatic and environmental factor	0.73	9	I	0.70	10	I	0.71	12	I
Delayed response to repairs	0.67	14	I	0.68	13	I	0.68	13	I
Lack of skilled maintenance personnel	0.67	14	I	0.69	12	I	0.68	13	I
Overcrowding and capacity overload	0.68	13	I	0.68	13	I	0.68	13	I

NOTE: EI = Extremely Influential, I = Influential, ME = Moderately Influential.
Source: Authors (2026).

The findings show that overall, budget constraints rank highest (RII=0.79, Extremely Influential), followed by inadequate monitoring systems (RII=0.77, Extremely Influential), spare parts availability and poor maintenance documentation (both RII=0.76, Extremely Influential), implying funding, oversight, and logistics as primary drivers of system failures in the students' hostels. Poor installation quality and inadequate preventive maintenance tie for fifth overall (both RII=0.75, Influential), highlighting initial workmanship and routine upkeep as key influences. Excessive user load/misuse, poor communication channels, lack of user education, and design inadequacy tie for seventh overall (all RII=0.73, Influential), pointing to user behavior and design flaws as moderate contributors. The findings also show that building age/degradation ranks eleventh overall (RII=0.72, Influential), climatic/environmental factors twelfth (RII=0.71, Influential), and delayed repairs, lack of skilled personnel, and overcrowding tie for thirteenth (all RII=0.68, Influential), suggesting these as lower but still influential factors.

Among maintenance staff (N=27), poor installation quality and poor communication tie for first (both RII=0.78, Extremely Influential), inadequate preventive maintenance third (RII=0.77, Extremely Influential), and budget constraints/excessive user load/inadequate monitoring tie for fourth (all RII=0.76, Extremely Influential), emphasizing technical and user issues. Spare parts availability/inadequate documentation tie for seventh among staff (both RII=0.75, Influential), climatic factors/lack of user education ninth (both RII=0.73, Influential), building age/design inadequacy eleventh (both RII 0.70, I), overcrowding thirteenth (RII=0.68, Influential), and delayed repairs/lack of skilled personnel fourteenth (both RII=0.67, Influential). Then, among porters (N=29), budget constraints rank first (RII=0.83, Extremely Influential), inadequate monitoring second (RII=0.78, Extremely Influential), and spare parts availability/inadequate documentation/design inadequacy tie for third (all RII=0.77, Extremely Influential), stressing financial and logistical barriers.

Building age/poor installation tie for sixth among porters (both RII=0.73, Influential), excessive user load tenth (RII=0.70, Influential), climatic factors tenth (RII=0.70, Influential), lack of user education/inadequate preventive maintenance eighth (both RII=0.72, Influential), and delayed repairs/lack of skilled personnel/overcrowding/poor communication tie for thirteenth (all RII=0.68, Influential). The Table shows that comparing groups, porters rank budget constraints higher (1st, RII 0.83 EI) than staff (4th, RII 0.76 EI), while staff prioritize poor installation (1st, RII=0.78 Extremely Influential) over porters (6th, RII=0.73 Influential), revealing role-specific emphases on funding vs. technical quality. Overall convergence on top factors like budget and monitoring suggests targeted interventions, with divergences (e.g., porters higher on design inadequacy, staff on user misuse) informing role-tailored strategies for hostel maintenance.

IV.4 TEST OF HYPOTHESIS

Two hypotheses were formulated for the study as follows:

Ho1: There is no statistical difference in the perception of maintenance staff and porters on the factors influencing failure of critical building systems in students' hostels.

Ho 2: There is no statistical difference in the factors influencing failure of critical building systems in students' hostels across the various types of institutions.

Mann-Whitney U test and Kruskal-Wallis test were used to analyze the hypotheses respectively. The Mann-Whitney U test evaluated the null hypothesis that there is no statistical difference in the perception of maintenance staff and porters on the factors influencing failure of building systems and components in students' hostels while the Kruskal-Wallis test evaluated the null hypothesis that there is no statistical difference in the factors influencing failure of building systems and components in students' hostels across the

various types of institutions. Table 5 presents the result for the first hypothesis indicating the mean rank scores, U-values, and *p*-values for both categories of respondents.

Table 5: Mann-Whitney U Test of difference in the perception of maintenance staff and porters on the factors influencing failure of critical building systems in students' hostels.

Factors influencing systems failure	Category of Respondents		U	Z	*p*	Remark
	Maintenance staff N=27	Porter N=29				
	Mean Rank (MR)	Mean Rank (MR)				
Budget constraints	25.43	31.36	308.500	-1.472	0.14	NS
Spare parts availability	26.80	30.09	345.500	-0.978	0.33	NS
Delayed response to repairs	28.13	28.84	381.500	-0.187	0.85	NS
Lack of skilled maintenance personnel	27.83	29.12	373.500	-0.322	0.75	NS
Overcrowding and capacity overload	28.67	28.34	387.000	-0.078	0.94	NS
Building age and degradation	26.83	30.05	346.500	-0.807	0.42	NS
Climatic and environmental factor	29.28	27.78	370.500	-0.370	0.71	NS
Poor installation quality	30.63	26.52	334.000	-1.015	0.31	NS
Excessive user load and misuse	31.02	26.16	323.500	-1.201	0.23	NS
Poor communication channels	34.44	22.97	231.000	-2.986	0.00	S
Lack of user education	28.81	28.21	383.000	-0.159	0.87	NS
Poor maintenance documentation	27.69	29.26	369.500	-0.389	0.70	NS
Design adequacy	24.93	31.83	295.000	-1.732	0.08	NS
Inadequate preventive maintenance	30.56	26.59	336.000	-0.990	0.32	NS
Inadequate monitoring systems	27.63	29.31	368.000	-0.440	0.66	NS

Significant level set at $p \leq 0.05$.
Source: Authors (2026).

The result in Table 5 shows that there is no significant difference in the perception of maintenance staff and porters on factors influencing failure of systems in students' hostels except for poor communication (p=0.00), hence, the null hypothesis was accepted for the factors the factors influencing failure of systems in students' hostels except for poor communication channels, for which the null hypothesis H_0 was rejected. Table 6 presents the result for the second hypothesis indicating the mean rank scores, degree of freedom (df), and p-values.

Table 6: Kruskal-Wallis Test of difference in the factors influencing failure of critical building systems in students' hostels across various types of institutions.

Factors influencing systems failure	FU (N=21)	FP (N=10)	PU (N=15)	FCE (N=10)	H (x^2)	df	*p*	Remark
	Mean Rank	Mean Rank	Mean Rank	Mean Rank				
Budget constraints	34.64	37.30	20.23	19.20	15.195	3	0.00	S
Spare parts availability	30.90	23.20	28.60	28.60	2.543	3	0.47	NS
Delayed response to repairs	29.88	32.40	27.17	23.70	2.199	3	0.53	NS
Lack of skilled maintenance personnel	28.31	24.45	28.40	33.10	1.682	3	0.64	NS
Overcrowding and capacity overload	30.29	34.55	28.47	18.75	5.883	3	0.12	NS
Building age and degradation	26.07	35.90	25.03	31.40	4.208	3	0.24	NS
Climatic and environmental factor	25.31	29.60	32.10	28.70	1.822	3	0.61	NS
Poor installation quality	23.74	33.65	31.70	28.55	3.899	3	0.27	NS
Excessive user load and misuse	26.07	26.35	35.37	25.45	4.234	3	0.24	NS
Poor communication channels	27.14	25.90	32.53	27.90	1.712	3	0.63	NS
Lack of user education	25.55	30.10	32.43	27.20	2.229	3	0.53	NS
Poor maintenance documentation	30.17	30.75	27.33	24.50	1.265	3	0.74	NS
Design adequacy	26.95	27.90	29.10	31.45	0.659	3	0.88	NS
Inadequate preventive maintenance	33.14	27.75	22.53	28.45	4.417	3	0.22	NS
Inadequate monitoring systems	29.88	27.70	27.20	28.35	0.353	3	0.95	NS

NOTE: FU is Federal University, FP is Federal Polytechnic, PU is Private University, FCE is Federal College of Education, Significant level is at $p \leq 0.05$.
Source: Authors (2026).

The result in Table 6 shows that there is no statistical difference in the factors influencing failure of building systems in students' hostels across the various types of institutions except for budget constraints (p=0.002) hence, the null hypothesis $H_0 2$ for all the factors was accepted, except for budget constraints for which the null hypothesis $H_0 2$ was rejected. The result suggests that the Federal University and Federal Polytechnic experiences significantly higher budget constraints with Mean Ranks of 34.64 and 37.30 respectively as compared to the Private University and the Federal College of Education with Mean Ranks of 20.23 and 19.20 respectively.

IV.5 DISCUSSION OF FINDINGS

The results from this study provide valuable insights into the reliability of critical building systems in student hostels within Lagos Metropolis, highlighting variations in Mean Time Between Failures (MTBF) across different systems and institution categories.

From the analysis of data collected from maintenance staff and porters working in the students' hostels of the surveyed institutions, the study quantifies system performance, identifies key influencing factors, and tests for differences in responses through statistical methods.

The findings of the study not only address the research objective of measuring average MTBF but also reveal institutional disparities and role-specific perceptions, contributing to a deeper understanding of maintenance challenges in educational infrastructure. The discussion is structured to explore the top five systems with critical MTBF values, linking empirical data to existing literature, and examining implications for practice. Among the critical building systems evaluated, foundation of the hostel buildings exhibits the highest MTBF values across the institutions, indicating high reliability and infrequent failures. This aligns with literature emphasizing the durability of foundations in most buildings, which are designed to withstand long-term loads and are less susceptible to daily wear and tear compared to mechanical systems [21]. In the context of students' hostel buildings in Lagos, where environmental factors like soil instability and high humidity could accelerate degradation, the high MTBF suggests appropriate choice of foundation design and generally, high quality of construction work when the hostel buildings were being constructed. Roofs, staircase and walls equally demonstrated high reliability with their respective high MTBF values, reflecting their robust performance in supporting the overall building framework.

This result concurs with [18] that structural components benefit from inherent resilience when properly maintained, reducing the frequency of failures [21]. It is also important to note that in student hostels, where occupancy is high and usage intense, this MTBF particularly in staircase components may be attributed to periodic maintenance inspections that prevent minor issues from escalating. The findings show that wiring systems, lighting systems and water pipes are the most frequently failing systems in students' hostels with relatively low MTBF values. Wiring systems failure in public buildings often arises from cable overheating and arcing fault as a result of overloaded circuits, loose connections, or improper cable use. This aligns with the position of [7,21] that public buildings are often challenged by maintenance demands relating to electrical systems. Furthermore, the findings revealed frequent light bulbs burn out in the students' hostels which could be due to poor quality of materials and parts or systemic electrical issues such as high voltage, poor or loose wiring, and voltage fluctuations. Other factors that could influence rapid failure of light bulbs include overheating from poor ventilation in the enclosed fixtures, incompatible switches and ratings, and physical damage from vibration or improper installation.

These findings highlight the need for standardized protocols in the procurement of quality materials and parts that could enhance systems and components longevity, particularly in resource-limited public settings. Water pipes failure, especially frequent plumbing pipes leakages in student hostels are often as a result of high-intensity use, poor maintenance, and aging infrastructure. This finding supports the position of [20, 21, 22] that plumbing systems in public buildings often suffer frequent defects including corrosion and blockages, but regular maintenance can improve their availability and extend operational periods [20]. Conclusively, the high MTBF values for structural elements such as foundations, walls and staircases illustrate areas of relative strength in Lagos hostel infrastructure while systems such as wiring systems, lighting bulbs and water pipes are exposed as being highly vulnerable to failure underscoring the need for improved attention, institutional funding and proactive maintenance approaches at improving their reliability. These findings portray the study's contribution to bridging research gaps in MTBF applications for educational facilities, advocating for targeted interventions to elevate system reliability and ultimately enhance student comfort and safety.

V. CONCLUSIONS

This study has illuminated the reliability of critical building systems in student hostels across Lagos Metropolis through MTBF analysis. The study concludes that structural and foundational elements are the most reliable with rare occurrence of failures in students' hostels while wiring systems, lighting bulbs and water pipes are frequently failing systems with relatively low reliability. The study also concludes that the reliability of systems in students' hostels are largely influenced by budget constraints, inadequate monitoring systems, spare parts unavailability and poor maintenance documentation. By quantifying failures and identifying influencing factors, the study addresses gaps in maintenance performance metrics for Nigerian educational infrastructure. To this end, the study recommends increasing budgetary allocations for operations and maintenance of infrastructure in tertiary institutions to enable preventive maintenance and technology driven maintenance implementation, such as computerized maintenance management systems (CMMS). It also recommends the implementation of regular staff training programs to improve skills in failure prediction and repair; development of standardized reporting protocols to enhance data accuracy and monitoring; promoting user education campaigns to reduce misuse and overload; and conducting longitudinal studies integrating MTTR for holistic key performance indicator frameworks. These measures can foster sustainable hostel management, boosting students' well-being and academic success.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Olajide Julius Faremi, Damilola Abdulrasaq Lukman and Dele Roger Simeon.
Methodology: Olajide Julius Faremi, Damilola Abdulrasaq Lukman and Dele Roger Simeon.
Investigation: Olajide Julius Faremi, Damilola Abdulrasaq Lukman.
Discussion of results: Olajide Julius Faremi, Damilola Abdulrasaq Lukman and Dele Roger Simeon.
Writing – Original Draft: Olajide Julius Faremi, Damilola Abdulrasaq Lukman and Dele Roger Simeon.
Writing – Review and Editing: Olajide Julius Faremi, Damilola Abdulrasaq Lukman and Dele Roger Simeon.
Resources: Olajide Julius Faremi, Damilola Abdulrasaq Lukman.
Supervision: Olajide Julius Faremi.
Approval of the final text: Olajide Julius Faremi, Damilola Abdulrasaq Lukman and Dele Roger Simeon.

VII. ACKNOWLEDGMENTS

The authors would like to offer their heartfelt appreciation to all respondents who generously contributed of their time and insightful thoughts during the data collection process. Their participation, openness, and insightful comments were critical to the study's success.

VIII. REFERENCES

[1] I. Idowu, B. E. Dickson, and L. Idowu, "Qualitative Analysis of the Relationship between Hostel Conditions and Academic Performance of University Students in South Western Nigeria," *Int. J. Contemp. Appl. Res.*, vol. 10, no. 2, pp. 2308–1365, 2023.

[2] D. N. Mbazor, "Assessment of the influence of on-campus housing quality and facilities on students' academic performance at the Federal University of Technology, Akure, Nigeria," *Baltic J. Real Estate Econ. Constr. Manag.*, vol. 9, pp. 14–32, 2021, doi: 10.2478/bjreecm-2021-0002.

[3] Z. B. Fatima, C. Belkacemic, and L. Zemmouchi-Ghomari, "Essential and new maintenance KPIs explained," *Int. J. Educ. Manag. Eng.*, vol. 12, no. 6, pp. 11–20, 2022, doi: 10.5815/ijeme.2022.06.02.

[4] Y. A. Wahab and A. S. H. Basari, "Analysis of down time and reliability estimation in hostel building maintenance – A case study," *Middle-East J. Sci. Res.*, 2013.

[5] I. C. Osuizugbo and B. O. Olusanya, "Maintenance management practices and factors affecting maintenance of private university buildings in Ogun State Nigeria," *Civil Sustain. Urban Eng.*, vol. 3, no. 2, pp. 112–122, 2023, doi: 10.53623/csue.v3i2.258.

[6] N. A. Sanusi, "A study on the building maintenance practices in students' hostels at public universities," *J. Social Sci. Humanit.*, vol. 16, no. 3, pp. 1–17, 2019.

[7] R. C. Ebele and W. Ebiwari, "Developing maintenance management supportive framework as measures for effective maintenance management practices in selected Nigerian universities," *Eur. J. Bus. Manag. Res.*, vol. 7, no. 1, pp. 147–152, 2022, doi: 10.24018/ejbmr.2022.7.1.1190.

[8] P. Olusegun, A. Kehinde, and A. Muhydeen, "Students' assessment of hostel facilities in The Polytechnic Ibadan, Nigeria: Realities and challenges," *J. Educ. Pract.*, vol. 10, no. 30, pp. 83–90, 2019, doi: 10.7176/jep/10-30-10.

[9] A. O. Oluwatobi, V. A. Alli-Johnson, C. A. Ayedun, and O. A. Akinjare, "Assessment of the effectiveness of maintenance management systems in delivering quality maintenance services in higher institutions," *J. Phys.: Conf. Ser.*, vol. 1299, no. 1, 2019, doi: 10.1088/1742-6596/1299/1/012015.

[10] B. Usman, "Evaluation on students hostel accommodation in tertiary institutions: A case study of some selected institutions in Borno State-Nigeria," *Int. J. Sustain. Dev.*, vol. 6, no. 4, pp. 1–8, 2021. [Online]. Available: https://www.arcnjournals.org/images/NIRA-USD-6-4-1.pdf

[11] S. Ferreira, L. Martins, F. J. G. Silva, R. B. Casais, R. D. S. G. Campilho, and J. C. Sá, "A novel approach to improve maintenance operations," *Procedia Manuf.*, vol. 51, pp. 1531–1537, 2020, doi: 10.1016/j.promfg.2020.10.213.

[12] L. A. Ndulue and F. Ifeanyiemoh, "Assessment of factors affecting effective hostel maintenance in Federal Polytechnic Oko, Anambra State Nigeria," *Int. J. Civil Eng., Constr. Estate Manag.*, vol. 9, no. 1, pp. 21–42, 2021.

[13] M. Jasiulewicz-Kaczmarek, S. Legutko, and P. Kluk, "Maintenance 4.0 technologies – new opportunities for sustainability driven maintenance," *Manag. Prod. Eng. Rev.*, vol. 11, no. 2, pp. 74–87, 2020, doi: 10.24425/mper.2020.133730.

[14] A. King, S. M. M., and P. Ugochukwu, "Evaluation of factors affecting student hostel maintenance in a tertiary institution in Kaduna South, Kaduna State, Nigeria," *J. Appl. Sci. Environ. Manag.*, vol. 28, no. 8, pp. 2319–2323, 2024.

[15] Y. A. Wahab and A. S. H. Basari, "Maintenance management practices for hostel facility: A case study," *J. Kejuruteraan Dan Sains Kesihatan J. Eng. Health Sci.*, vol. 3, pp. 119–128, 2019.

[16] J. T. Selvik and E. P. Ford, "Down time terms and information used for assessment of equipment reliability and maintenance performance," *IntechOpen*, vol. 11, pp. 13, 2017.

[17] W. Torell and V. Avelar, "Mean time between failure: Explanation and standards," *Power*, vol. 78, pp. 1–10, 2004. [Online]. Available: http://support.casit.net/Portals/0/NTForums_Attach/VAVR-5WGTSB_R0_EN.pdf

[18] A. Singh, S. Adachi, and M. Inouye, "Quality control analysis of downtime and time to repair for water supply pipes," *Built Environ. Proj. Asset Manag.*, vol. 1, no. 1, pp. 75–90, 2011, doi: 10.1108/20441241111143795.

[19] A. Żyluk, M. Zieja, N. Grzesik, J. Tomaszewska, G. Kozłowski, and M. Jasztal, "Implementation of the mean time to failure indicator in the control of the logistical support of the operation process," *Appl. Sci. (Switz.)*, vol. 13, no. 7, 2023, doi: 10.3390/app13074608.

[20] T. Williams, "5 common plumbing problems and how to diagnose them," *Architectural Digest*, Jun. 16, 2022. [Online]. Available: https://www.architecturaldigest.com/story/plumbing-problems

[21] S. Yacob, A. S. Ali, and C. P. Au-Yong, "Establishing relationship between factors affecting building defects and building condition," *J. Surveying, Constr. Prop.*, vol. 10, no. 1, pp. 31–41, 2019, doi: 10.22452/jscp.vol10no1.3.

[22] M. M. Salihu, M. Hassan, and N. B. Polytechnic, "Student satisfaction with facility provision and quality in Nuhu Bamalli Polytechnic Zaria, Kaduna State," vol. 1, no. 1, pp. 9–20, 2023.

[23] E. O. Lucky and S. A. Iyanda, "Evaluation of users' satisfaction with hostel management facilities and services in tertiary institution in Minna," *Afr. Scholar Publ.*, Sep. 2022. [Online]. Available: http://www.africanscholarpublications.com/wp-content/uploads/2023/01/AJBEGR_Dec_2022_Vol27_No4-21.pdf

[24] Y. A. Wahab, "Cost-downtime model technique," vol. 7, no. 1, pp. 55–69, 2023.

[25] M. Sa'ad, H. Babayo, B. Ishiyaku, and A. Sa'ad, "Students' satisfaction with essential hostel services provision and academic performance," *J. Energy Environ.*, 2020.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 366-374. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3000

RESEARCH ARTICLE OPEN ACCESS

FORMABILITY AND SIZE EFFECTS IN MICROFORMING: A REVIEW OF MATERIALS, METHODS, AND APPLICATIONS

Gopi Durgam*[1], and Kishor K Dhande[2]

[1, 2]Department of Mechanical Engineering, Dr. D Y Patil Institute of Technology, Pimpri, Pune 411018, Maharashtra, India

[1]http://orcid.org/0009-0002-7787-2934, [2]http://orcid.org/0000-0002-2483-3766

Email: *gopidurgam76@gmail.com, kishor.dhande@dypvp.edu.in

ARTICLE INFO

Article History
Received: November 24, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Example one,
Example two,
Example three,
Example four,
Example five.

ABSTRACT

Micro-manufacturing has drawn significant global attention regarding manufacturing technologies and procedures. Micro-forming is a highly popular method for micro-manufacturing. Microforming is a process of shaping parts and objects through mechanical deformation, which relies on the properties of materials. A significant amount of attention has been directed towards micro-forming, specifically the deep drawing process, due to its capability to generate a wide range of products, especially in its traditional macro-process. This technique is employed to fabricate the most commonplace objects. The discussion then transitions to the current areas of research that are generating the greatest discussion. Due attentiveness has shown a gap in micro-forming technological improvement that must be filled before it can satisfy the unique demands of industrial applications. This is of utmost importance when it comes to making top-notch end products. Precision-forming machines can manufacture such goods using cutting-edge tooling technology, optimal operating settings, and improved material handling techniques.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

In addition to being an essential part of modern manufacturing, metal shaping is a sizable business in and of itself. Hundreds of millions of tons of metal go through metal-forming processes every year, all around the globe. In most industrialized countries, the metal-forming sector accounts for around 15% to 20% of GDP. It also serves a social purpose by giving many people a chance to earn an income. The metal forming industry is all about mass manufacturing of both partially finished and fully assembled products. Significant advantages for large-scale R&D may be found in this area, since even little savings per ton can add up to big bucks. To enhance the accuracy of the Forming Limit Diagram, surface strain data were further included.

A circular grid made of foil is essential for measuring surface strain [1]. The forming limit diagram (FLD) is a valuable tool for assessing the formability of a metal cost-effectively [2]. Conducting tests utilizing a hemispherical punch-stretch or a Marciniak cup might result in forming limit diagrams (FLDs). Significant amounts of time and effort were dedicated to developing these experimental approaches. Increasing the efficiency of FLDs is accomplished using analytical approaches. Regrettably, analytical methodologies do not possess the capability to foresee failure precisely. A forming limit diagram, also known as an FLD, is a graphical depiction that depicts the main strains (e1) at the point when localized necking occurs, for each conceivable value of the minor strain (e2) [3].

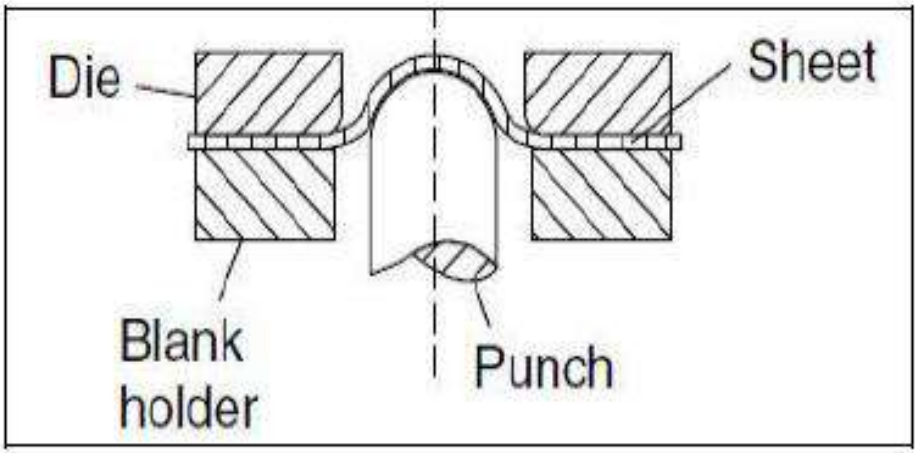


Figure 1: Sheet metal forming.
Source: [4].

Steel facilities, such as factories and plants, are responsible for manufacturing sheet metal, which is used in manufacturing vehicles and their components. Various companies use sheet metal as the primary material for manufacturing various home and commercial goods. Steel makers generate billets that re-rolling mills then shape into various components, including angles, channels, and bars. Incremental forming technique for manufacturing micro-parts that eliminates the need for dies or backing plates [5],[6]. According to [7], this technology may simplify tool manufacture, but it is unsuitable for large-scale production of small components. The micro-sheet metal forming industry utilizes laser forming, a die-less technique, as shown in Figure 1. According to [8] suggest that laser-assisted bending might be used to manage spring-back. The by [9] comprehensively described the process parameters and material requirements involved in laser bending.

The potential use of laser technology in micro-deep drawing methods is now under consideration. In [10] found spark formation more effective than laser forming. Spark formation has several benefits, including its versatility in working with diverse materials, cost-effectiveness, compactness in fabrication equipment, and low energy use.The challenges of microforming, particularly with ultrathin sheets of less than 100 microns, are well recognized in the field, where material deformation can lead to defects such as necking, cracks and wrinkles. As the demand for miniaturization increases in various engineering applications, the need for precise microforming processes also increases. To address this need, the article presents a new tool specifically designed to investigate the formability of thin titanium grade II (TiG-II) sheets, focusing on microstructural behavior. It provides a basis for future research, particularly in applying grade II titanium in micro-sized biomedical components [11].

The study focuses on developing forming limit curves (FLC) for TaRO5200 with a sheet thickness of 80 microns, a fundamental tool for evaluating the formability of microscale materials. This variation suggests that numerical simulations provide a conservative design approach, 15% safer than experimental results. This research's novelty lies in applying the FLC methodology to TaRO5200 and in the detailed microstructural study carried out on the specimens before and after the forming tests. This microstructural analysis provides valuable insights into the material behavior and underlying mechanisms during the formation process, contributing to a deeper understanding of the physics governing the deformation of TaRO5200. According to [12] findings provide a basis for further exploration of biomaterial microforming processes, particularly in biomedical device manufacturing. The research used Simufact Forming V15 software to generate the FLC numerically, and the results were subsequently compared with those obtained from the Nakajima test.

The comparison demonstrated good agreement between the experimental and numerical approaches, and the FLCs produced by numerical simulation were 5% to 12% lower than the experimental results, indicating a safer design margin in the simulations. Additionally, microstructural studies were performed on SS316L sheets before and after forming tests, providing a deeper understanding of the material's behavior during deformation. This comprehensive study provides valuable information on the microforming of SS316L, particularly its application as a biomaterial, and highlights the importance of accurate FLC determination to ensure the reliability of microformed components [13].The study of [12] focuses on the formability evaluation of Zr702 thin sheets, emphasising the development of forming limit curves, a critical tool in evaluating material formability.

The research uses the Nakajima test, performed according to ASTM 2218-14 standards, to experimentally determine FLCs for Zr702 with a sheet thickness of 100 microns. The results of these experimental FLCs were then compared with those obtained through numerical simulations, revealing good agreement between the two methods. In addition to the FLC analysis, a detailed microstructural study of the test specimens was performed before and after the forming process, which provided valuable information on the behavior of the material during deformation. This research contributes significantly to the understanding of the formability and microstructural characteristics of Zr702 in microforming applications, offering a basis for its broader adoption in biomedical fields [14]. Regarding mass manufacturing, none of the technologies mentioned can hold their own. Rather, they should only be used for prototypes or low-volume manufacturing.

Furthermore, certain requirements linked to material characteristics limit the variety of materials used. Remember that the methods mentioned here are complementary, not competing, regarding micro-metal formation. Mechanical micro-sheet metal forming remains the most efficient option for producing micro-components. After 10 years of extensive study and development, the area of mechanical micro-sheet metal forming has yielded a few results. To detect the gaps in the growth of micro-sheet metal, it is necessary to have a mapping that can provide a full picture. This study thoroughly examines the researchers' aims and the necessary conditions for advancing mechanical micro-sheet metal forming technology. To ensure a thorough analysis, the researchers have classified their results based on the characteristics of micro-sheet metal-forming systems, such as the material, process, tool, and machine parameters.

II. THEORETICAL REFERENCE

II.1 SUBTITLE

II.1.1 Materials and Methods

As a result of the shrinking of goods, which allows micro-geometry to carry out product tasks, the present tendency in numerous engineering disciplines is to reduce the size of products. Waste, expenditure, weight, and volume are all areas that will be reduced with this strategy [15]. The microfabrication method may take advantage of metals' outstanding mechanical and functional properties in manufacturing tiny components. This cutting-edge technology is ideal for use in manufacturing since it can be scaled down to the micron size range thanks to this process. According to [16], the market for microelectromechanical systems (MEMS) in micro-manufacturing is anticipated to rise at a compound annual growth rate (CAGR) of 7.5% for medical applications, 43.6% for telecom applications, and 11.0% for industrial applications between the years 2015 and 2024. There has been a significant increase in the demand for small-sized goods because of the growing trend toward downsizing in several different sectors. The medical, precision, communication, micro-electromechanical, and microfluidic equipment industries are especially affected by this phenomenon.

Over the last two decades, there has been significant development in the field of metal micro-forming technology. A variety of research and development activities are included in the study [13]. These activities include surface engineering for die and tooling, micro-hydroforming, embossing, micro-bulk forming, micro-forming with laser help, micro-sintering, and micro-sheet metal forming. The size affected the flow, necessitating an investigation of the deformation behaviours induced by the flow. In this work, we conducted experimental trials to examine the influence of many important factors, including geometry and grain sizes, on the extent of flow-induced defects that arise during the micro-forming process of pure copper. These investigations aimed to ascertain the impact of size on flow-induced flaws [17], [18]. Flow faults in the micro-extrusion process occur due to uneven material flow. This condition is analogous to the one that arises in the macro-forming process.

II.2 SIZE EFFECTS IN MICRO FORMING PROCESSES

An investigation was conducted on the influence of size effects on multi-phase materials, which are affected by the specimen's geometry and grain size. The research demonstrated a reduction in ductility as the size of the specimens was scaled to the millimetre scale range. Moreover, the research highlighted the importance and relevance of having a detailed and intricate surface. It is crucial to consider the impact of size-related factors, such as specimen shape and grain size, especially in materials with several phases, to enhance material characteristics and performance [19],[20]. During the downsizing and micro-component manufacturing process, the usual principles that apply to larger objects are sometimes not applicable.

The side effects, which are caused by the decreased number of grains in the deformation zone, significantly impact formability and might result in fractures during micro forming operations. This emphasizes the need of meticulously considering the impacts of size in micro-forming processes [21]. The work conducted by [22] reveals that the forming limit of copper foils is positively correlated with an increased number of grains. They tested copper foils with 25 and 500mm diameters by subjecting them to bulging, tensile, and bending loads. A spring-back effect, brought on by the smaller particle size and the more significant proportion of surface grains, diminished the hardening capacity of the material, because of that decrease in size, the flow stress changed [23-26].

II.3 DIMENSION EFFECT ON FLOW STRESS

In microforming, the flow stress is significantly influenced by side effects, such as grain size and specimen size [27]. The observation made by [28] stated that, Micro-forming processes need careful analysis of factors, such as the effects of particle size on bulk materials. The behaviour of materials changes as the scale changes, highlighting the need to conduct investigations and analysis. The uniaxial tensile testing trials revealed an inverse relationship between flow stress and strip thickness, indicating that the reduction in flow stress is due to size or scale effects. A surface model was used to clarify these data. The research of [29] dives into the influence of size effects on flow stress inside metal micro-forming techniques, conducting micro tensile tests on pure copper foil to assess the combined impact of workpiece geometry and grain size.

A unique composite material model is built to quantify the strengthening reaction of individual grains included inside a polycrystalline material. Studies have also shown that the flow stress variations are closely related to the ratio of thickness to grain size (T/D) and the interaction between workpiece geometry and grain size [30-32]. A constitutive model was established to support multi-stage microforming. Additionally, incorporating a hybrid material model considering grain heterogeneity, sample thickness, and dislocation structures has proven effective in describing the hardening behavior of grains in microforming processes. These findings emphasize the critical role of side effects and material characteristics in determining flow stress during microforming.

II.4 EFFECTS OF TEMPERATURE ON MICRO FORMING

Temperature control significantly impacts material behavior in microforming processes, influencing material softening and ductility. Higher temperatures enhance ductility, reduce flow stress, and prevent failure. Controlled heating promotes homogeneity, minimizes surface defects, and influences grain size. However, temperature management must balance these benefits with thermal expansion effects for dimensional accuracy. Also, increasing temperature in microforming enhances formability by activating more slip systems, reducing anisotropic behavior, and improving forming homogeneity and reproducibility due to size effects compensation [33].Temperature plays a crucial role in micro-forming processes, influencing material behaviour, flow stress, surface roughness, and fracture mechanisms [34]. High temperatures increase the percentage of elongation in materials like stainless steel-304 foil, affecting tensile strength and fracture patterns. The goal of this study is to look at the mechanical features of a 200 μm thick stainless steel-304 foil by testing it at room temperature (RTT) and at high temperature (HTT).

It was found that the strain rate and temperature affect the flow stress behaviour, surface roughness, microstructural development, and fracture process. High-temperature deformation leads to recrystallization and grain growth, changing surface roughness and fracture mechanism. This analysis can guide the design of micro-sheet metal forming processes for producing micro-parts [35]. In hot embossing of PMMA for microchannel fabrication, temperature variations impact channel dimensions significantly, with optimal conditions identified at 150 °C and 295 kPa [36]. Additionally, in gas-assisted extrusion forming of plastic micro-tubes, higher gas temperatures reduce melt viscosity and enhance flow velocity [37]. An elevation in temperature during cold stamping techniques leads to modifications in the tool-sheet metal tribosystem. This modification can create issues with forming ability, as it affects the level of friction, lubricating properties, and the material's behaviour [38].

Understanding these temperature effects is crucial for optimizing micro-forming processes and ensuring the quality of micro parts. SODICK and WEDM software was used for precision forging micro gear moulds at different temperatures. The surface roughness of the moulds was less than 280 nanometers, and the moulds had a dimensional tolerance that was less than or equal to 2.5 µm. To achieve a surface roughness of 300 nm and a reduction in dimensional tolerance from -38.8 µm to +0.9 µm, the high-temperature forging technology effectively exhibited its outstanding precision [39]. Experimental findings reveal that temperature and grain size considerably impact the deformability of pure copper in micro forming [40]. In another study, Micro-deep drawing (MDD) is used in various fields, including micro-electromechanical systems, vehicle engineering, and chemical engineering. MDD tests on two-layer stainless steel-copper composite foils showed that complete circular cups cannot be formed due to poor formability. An optimal annealing temperature of 800 C was found for high surface quality, indicating the potential of MDD in these applications [41].

By increasing the temperature at which the deformation takes place, a higher number of slip systems within the polycrystalline metal are activated. This results in a shaping process that is more uniform and has improved consistency. In this study, a YAG laser beam is used to apply heat to the workpiece. Additionally, a finite element model (FEM) is developed to simulate the melting and cooling rate (MCWR) of pure copper. A numerical study of surface asperity, which is a sign of material heterogeneity at a small scale, is undertaken, and the results show a high agreement with the findings of the experiments [36]. Forming titanium at elevated temperatures improves its micro-forming behaviour, enhancing formability and forming limits compared to room temperature, as investigated in the study [42]. The work of by [43] presents a micro forming device with controlled heating up to 450°C for medical production, focusing on small, high-precision, low-weight, and high- strength components. The device analyzes magnesium alloy AZ31B flow stress and formability, forming dental components at varying temperatures.

III. MICRO DEEP DRAWING (MDD) PROCESS

Micro deep drawing, it is possible to fabricate hollow objects with slender walls that have a resemblance to miniature cups or boxes. The decrease in the coefficient of friction between the material and the tools due to using lubricant might impact the quality of goods generated in micro deep drawing at a small scale. The research conducted experiments on stainless steel 304 foils with different thicknesses to investigate the influence of blank holder forces and the T/D ratio on the limit draw ratio. The experiments used micro deep drawing and tensile testing. The results indicate that the LDR achieves stability when the T/D (Thickness-to-Diameter) ratio exceeds 10. This suggests enhanced formability and consistent behaviour in a deep drawing to produce micro sheets [44], as seen in Figure 3 and 4. Researchers Zhengyi Jiang and others have shown a micro deep drawing (MDD) simulation, a promising technique for mass-producing complex three-dimensional, tiny metal objects. A Voronoi blank model is constructed for microscale modeling.

This model enables the inclusion of material size effects and authentic microstructures. Simulation results show surface roughness significantly influences springback, drawability, and product quality, making the model accurate and beneficial for MDD process development [45]. Cooper foils are utilized in the microelectrical sector for deep drawing cup-like structures. One research explored micro deep sketching using T2 copper foils and three scale factors. Results revealed the effective formation of pieces with internal dimensions of 0.5, 1.0, and 2.0 mm. The maximum deep drawing force dropped with lower scale factors, and the maximum limit drawing ratio was 2.2 for the experiment with a 0.5 scale factor [46]. The process has also been used to create SUS304 circular cups, with simulation results showing good agreement with experimental data [47]. A novel technique using a floating ring as an assistant die has been proposed for micro-flexible deep drawing, with promising results in producing SS304 micro cups [48].

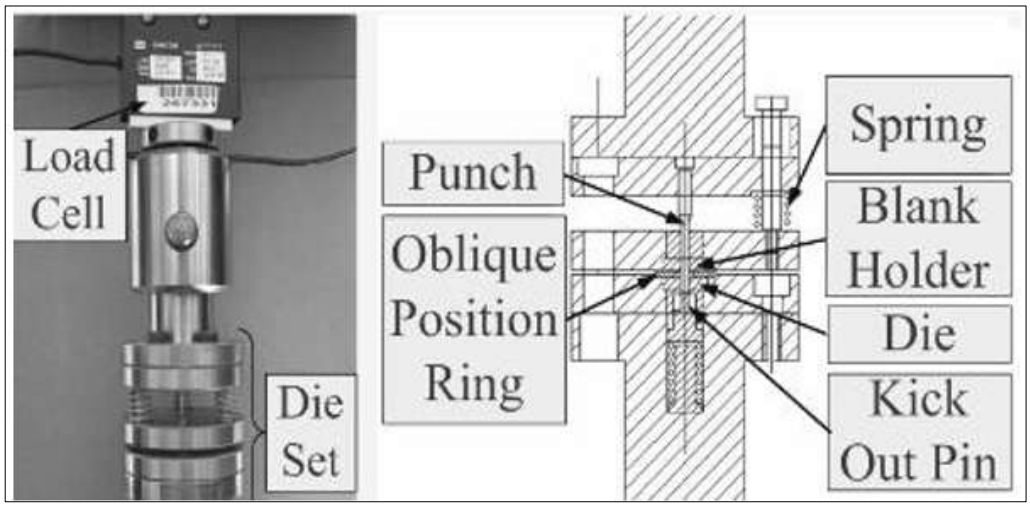


Figure 2: The Setup for Micro Deep Drawing Experiment.
Source: [47].

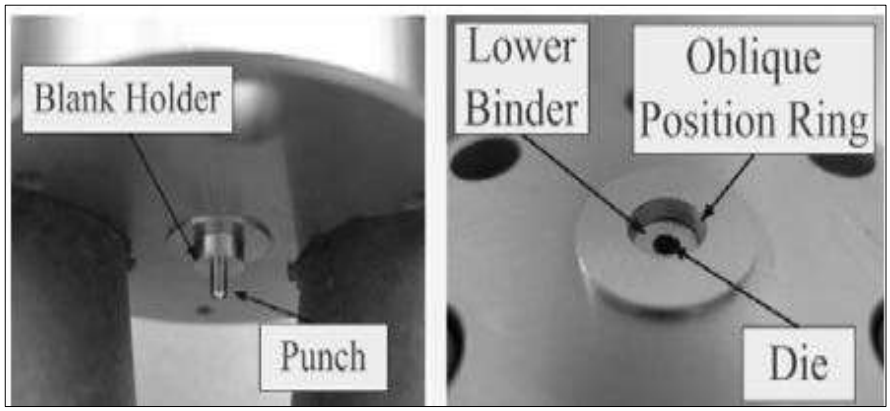


Figure 3: Micro Deep Draw Die Components.
Source: [47].

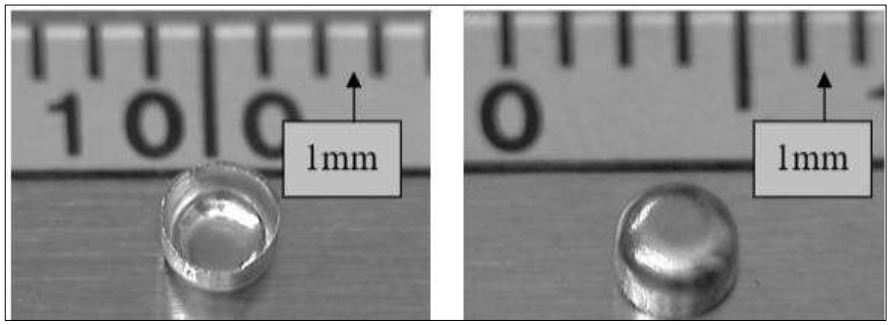


Figure 4: 50μm Thickness Cup (Drawing Ratio = 1.75).
Source: [47].

The study conducted by [49] examines the impact of forming velocity on the quality of micro cup-like products, specifically focusing on stainless steels with different thicknesses. The results indicate that the speed at which the forming process is carried out affects the profile's precision and the surface's quality. The resultant micro cup is rounder and more symmetrical when thicker blanks are used. Wrinkling becomes apparent when using thinner blanks, whereas the surface becomes smoother when the drawing velocity or blank thickness increases. The study by [50] focuses on the micro-deep drawing process of stainless steel 304 foils. The researchers used soft dies made of polyurethane rubber with different hardness levels. The process employs a mobile blank holder backed by a compressive spring.

There is an initial space between the blank holder and the adjustment ring, which controls the forming process against the force exerted by the spring. The calculations and testing examined key aspects such as the kind of rubber, the size of the die, and the scaling factor. These investigations confirmed that the approach can cost-effectively produce high-quality micro-cups. This study examines the micro-deep drawing technique of stainless steel 304 foils using soft dies made of polyurethane rubber with varying degrees of hardness. The process employs a movable blank holder backed by a compressive spring. There is an initial space between the blank holder and the adjustment ring, which controls the forming process against the force exerted by the spring. The models and testing investigated key aspects such as rubber type, die size, and scaling factor, and confirmed that the approach can cost-effectively produce high-quality micro-cups [51].

III.1 ANISOTROPY AND FORMABILITY

Before the 1980s, the yield standards for materials did not include the important anisotropy ratios, which made them less accurate. That's how things were until the 1980s. In 1948, Hills was the only thing that mattered in a few very stressful cases, and it was Hills that gave anisotropic materials the output they needed. Hill added orthotropic materials to the Von Mises criterion. Orthotropic materials are identified by their orthotropic symmetry [52]. According to [53] in his work had discussed the strain-hardening index (equation 1) wasn't the only thing that affected the height of the crack in the Olsen test. Because of this, the strain-rate sensitivity index must also be considered. Researchers found that the material would rise more when the strain-rate sensitivity index (m) was higher for a certain strain-hardening index (n). The sheet metal should be easier to shape now that this is clear [54]. Because of this, steel has a higher strain-rate sensitivity index than aluminum, which means that for a given number of n, it can be shaped better than aluminum.

$$m = \frac{\partial ln l\sigma}{\partial \dot{\varepsilon}} \tag{1}$$

Materials that are light, dense, strong, and easy to shape are now being sought after by the electronics and car industries around the world. Because of this, Tailor Welded Blanks have become more common because they can join different materials in thickness, properties, and surface conditions by shaping or pressing them together. When you use this method, the links between the parts are strong and will last long [55].

According to [56] did four tests on tailor-welded flats to see how easily they could be shaped using two different base materials. Aluminum-killed drawing quality (ASDQ) steel (1.8 mm) and high-strength low alloy steel (2.1 mm) were mixed to make the base. They came up with two reasons why it didn't work. If the main line of the extension was straight across from the weld, a kind 1 failure happened. When this kind of failure happened, the split depended on how well the material could bend and bend again. Furthermore, different sheet metal processes also form materials at very high temperatures, so it is crucial to determine the formability in these situations. Therefore, formability study of such novel approaches is also required to guarantee cost savings, minimal scrap generation, and the ability to apply the forming processes to various materials.

IV. CONCLUSIONS

This review aims to help the reader understand the significance of sheet metal forming techniques and how formability analysis plays a key role in them. The document briefly summarises the different tests used to obtain formability data that can be quantified and determine the forming limits beyond which failure occurs. Over the past 40 years, formability has advanced significantly, and research is still being done to find new materials, thorough testing procedures and tools, and mathematical analyses of the yielding criterion that consider a wide range of additional parameters and their effects on formability. Thus, this research investigates the fundamental aspects of formability and its significance in today's industry. Designing micro-components with high-volume production in mind is essential for micro-manufacturing. Building the failure limit diagram is now the focus of sheet metal research. This is accomplished using numerical analysis, experimental methods, finite element techniques, and sheet metal material management. This study highlights the need for more research into the interaction between materials and handling instruments and micro-forming processes (namely, deep drawing and bending operations).

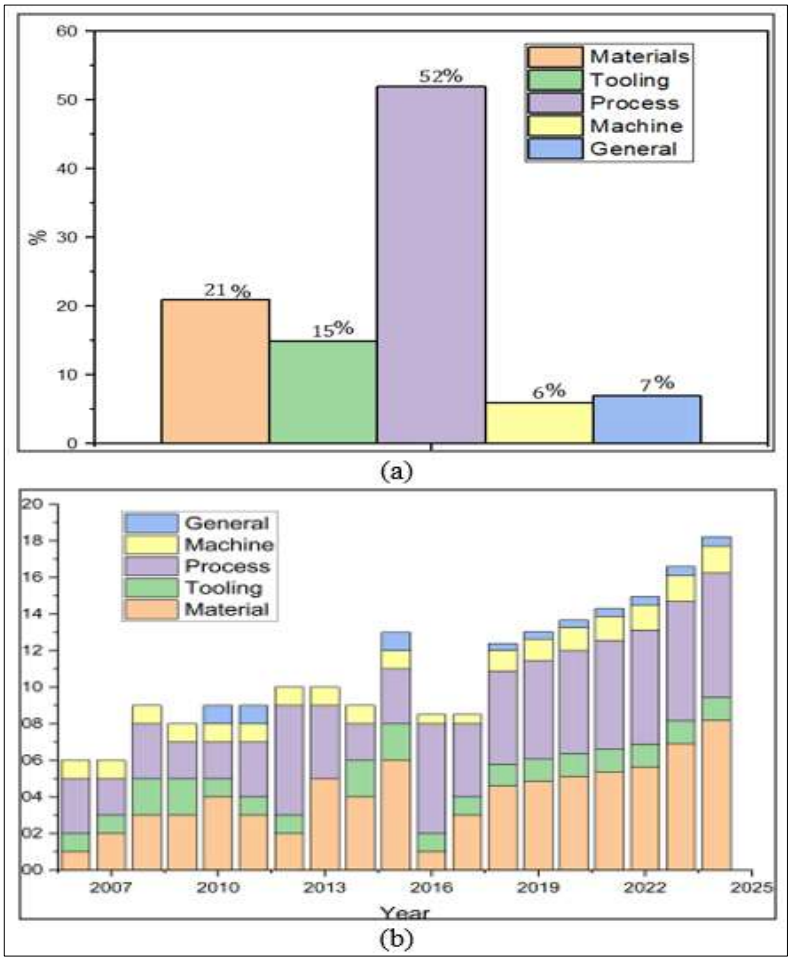


Figure 5: Research focuses on micro-sheet-metal-forming development, (a) in 2001–2017 (b) detailed for 2001–2017.
Source: Authors, (2026).

Figure 5 displays the results of the study's clustering, which shows that the focus was on improving micro-sheet metal-forming technology. Compared to other fields of study, process inquiry is intrinsically more interesting. According to the study's summary, 52% of the research carried out between 2001 and 2017 focused on improving process factors. Researchers are examining the behaviour of material deformation to improve micro-product quality. Nevertheless, many moving parts in manufacturing must be in perfect harmony for the final product to be high-quality. These include superior materials, cutting-edge tools, ideal working conditions, and effective material management. The use of a precision shaping machine might solve these problems. Consequently, more efforts are needed to enhance micro-forming applications in industry. Research product efficacy and process understanding might both be improved by industry collaboration. Many investigations on the micro-forming process have been carried out. However, more efforts are needed to rectify the shortfall in micro-forming technological improvement. The needs of industrial applications can only be satisfied by developing an integrated microforming system.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Gopi Durgam, K K Dhande.
Methodology: Gopi Durgam.
Investigation: Gopi Durgam.
Discussion of results: K K Dhande.
Writing – Original Draft: Gopi Durgam.
Writing – Review and Editing: K K Dhande.
Resources: K K Dhande.
Supervision: K K Dhande.
Approval of the final text: K K Dhande.

VI. REFERENCES

[1] G. Patel and G. Kakandikar, "Investigations on effect of thickness and rolling direction of thin metal foil on forming limit curves in microforming process," Modern Manufacturing Processes, pp. 145–155, 2020, doi: 10.1016/b978-0-12-819496-6.00007-5.

[2] U. V. M., "Comparative Studies on Formability Analysis in Metal Forming," International Journal of Research in Engineering and Technology, vol. 3, no. 1, pp. 196–200, 2014, doi: 10.15623/ijret.2014.0301031.

[3] V. R. Shinge and U. A. Dabade, "Experimental Investigation on Forming Limit Diagram of Mild Carbon Steel Sheet," Procedia Manufacturing, vol. 20, pp. 141–146, 2018, doi: 10.1016/j.promfg.2018.02.020.

[4] N. Tiwari, G. Kakandikar, and O. Kulkarni, "Micro Forming and its Application: A Critical Review," Journal of Engineering Research and Sciences, vol. 1, no. 3, pp. 126–132, 2022, doi: 10.55708/js0103013.

[5] T. Obikawa, S. Satou, and T. Hakutani, "Dieless incremental micro-forming of miniature shell objects of aluminum foils," International Journal of Machine Tools and Manufacture, vol. 49, no. 12–13, pp. 906–915, 2009, doi: 10.1016/j.ijmachtools.2009.07.001.

[6] Y. Saotome, K. Yasuda, and H. Kaga, "Microdeep drawability of very thin sheet steels," Journal of Materials Processing Technology, vol. 113, no. 1–3, pp. 641–647, 2001, doi: 10.1016/S0924-0136(01)00626-4.

[7] S. Deb, "Roll Forming in Sheet Metal Forming," Analysis and Optimization of Sheet Metal Forming Processes, pp. 1–20, 2024, doi: 10.1201/9781003441755-1.

[8] A. Gisario, M. Barletta, C. Conti, and S. Guarino, "Springback control in sheet metal bending by laser-assisted bending: Experimental analysis, empirical and neural network modelling," Optics and Lasers in Engineering, vol. 49, no. 12, pp. 1372–1383, 2011, doi: 10.1016/j.optlaseng.2011.07.010.

[9] F. Vollertsen, I. Komel, and R. Kals, "The laser bending of steel foils for microparts by the buckling mechanism—a model," Modelling and Simulation in Materials Science and Engineering, vol. 3, no. 1, pp. 107–119, 1995, doi: 10.1088/0965-0393/3/1/009.

[10] M. Otsu, T. Wada, and K. Osakada, "Micro-bending of Thin Spring by Laser Forming and Spark Forming," CIRP Annals, vol. 50, no. 1, pp. 141–144, 2001, doi: 10.1016/S0007-8506(07)62090-3.

[11] O. Kulkarni and G. Kakandikar, "Novel product design of tool for investigating formability with microstructural study of bio-material titanium grade-II thin foils," International Journal on Interactive Design and Manufacturing (IJIDeM), vol. 17, no. 5, pp. 2765–2775, 2022, doi: 10.1007/s12008-022-00903-3.

[12] O. Kulkarni and G. Kakandikar, "The Investigations on Formability of Tantalum RO5200 Thin Foils: Bio-Material," IOP Conference Series: Materials Science and Engineering, vol. 1284, no. 1, p. 012029, 2023, doi: 10.1088/1757-899X/1284/1/012029.

[13] M. N. Bogar, O. Kulkarni, and G. Kakandikar, "Micro Forming Studies of SS316L as Biomedical Application Material," Journal of Engineering Science and Technology Review, vol. 16, no. 4, pp. 133–141, 2023, doi: 10.25103/jestr.164.17.

[14] O. Kulkarni and G. Kakandikar, "Formability Assessment with Microstructural Investigations for Zirconium 702 Thin Foils: Bio-Material Applications," Advances in Materials and Processing Technologies, vol. 8, sup4, pp. 2367–2377, 2022, doi: 10.1080/2374068X.2022.2044131.

[15] A. Mashalkar, G. Kakandikar, and V. Nandedkar, "Micro-forming analysis of ultra-thin brass foil," Materials and Manufacturing Processes, vol. 34, no. 13, pp. 1509–1515, 2019, doi: 10.1080/10426914.2019.1655158.

[16] K. Manabe, "Metal Micro-Forming," Metals, vol. 10, no. 6, p. 813, 2020, doi: 10.3390/met10060813.

[17] J. L. Wang, M. W. Fu, and J. Q. Ran, "Analysis of size effect on flow-induced defect in micro-scaled forming process," The International Journal of Advanced Manufacturing Technology, vol. 73, no. 9–12, pp. 1475–1484, 2014, doi: 10.1007/s00170-014-5947-8.

[18] M. A. Musa, A. R. Razali, and N. I. Kasim, "Grain and Feature Size Effect on Material Behavior for Micro-Sheet-Forming," Applied Mechanics and Materials, vol. 680, pp. 77–80, 2014, doi: 10.4028/www.scientific.net/AMM.680.77
.
[19] K. Decamp et al., "Size and geometry effects on ductile rupture of notched bars in a C-Mn steel: experiments and modelling," International Journal of Fracture, vol. 88, pp. 1–18, 1997, doi: 10.1023/A:1007369510442.

[20] Z. T. Xu, L. F. Peng, M. W. Fu, and X. M. Lai, "Size effect affected formability of sheet metals in micro/meso scale plastic deformation: Experiment and modeling," International Journal of Plasticity, vol. 68, pp. 34–54, 2015, doi: 10.1016/j.ijplas.2014.11.002.

[21] T. Altan and A. E. Tekkaya, "Metal Forming Processes in Manufacturing," Sheet Metal Forming, pp. 1–4, 2012, doi: 10.31399/asm.tb.smff.t53400001.

[22] A. Diehl, U. Engel, and M. Geiger, "Influence of microstructure on the mechanical properties and the forming behaviour of very thin metal foils," The International Journal of Advanced Manufacturing Technology, vol. 47, no. 1–4, pp. 53–61, 2008, doi: 10.1007/s00170-008-1851-4.

[23] H. Hoffmann and S. Hong, “Tensile Test of very thin Sheet Metal and Determination of Flow Stress Considering the Scaling Effect,” CIRP Annals, vol. 55, no. 1, pp. 263–266, 2006, doi: 10.1016/S0007-8506(07)60412-0.

[24] L. V. Raulea, A. M. Goijaerts, L. E. Govaert, and F. P. T. Baaijens, “Size effects in the processing of thin metal sheets,” Journal of Materials Processing Technology, vol. 115, no. 1, pp. 44–48, 2001, doi: 10.1016/S0924-0136(01)00770-1.

[25] M. Karalar and M. Bayramoğlu, “Combined impacts of thickness and bending angle on springback of 1000DP steel sheets,” Ironmaking & Steelmaking, vol. 49, no. 7, pp. 693–698, 2022, doi: 10.1080/03019233.2022.2038010.

[26] J. Davy, “Miniaturisation among the Makah,” Worlds in Miniature, pp. 61–81, 2019, doi: 10.2307/j.ctvf3w24f.10.

[27] W. L. Chan and M. W. Fu, “Studies of the interactive effect of specimen and grain sizes on the plastic deformation behavior in microforming,” The International Journal of Advanced Manufacturing Technology, vol. 62, no. 9–12, pp. 989–1000, 2012, doi: 10.1007/s00170-011-3869-2.

[28] M. Singh, A. Hossain, and D. B. Wei, “A Hybrid Model for Studying the Size Effects on Flow Stress in Micro-Forming with the Consideration of Grain Hardening,” Key Engineering Materials, vol. 794, pp. 97–104, 2019, doi: 10.4028/www.scientific.net/KEM.794.97
.
[29] C. Wang et al., “Modeling of flow stress size effect based on variation of dislocation substructure in micro-tension of pure nickel,” Materials Research Express, vol. 4, no. 12, p. 126502, 2017, doi: 10.1088/2053-1591/aa9b66.

[30] R. Zhang, Z. Xu, L. Peng, X. Lai, and M. W. Fu, “Modelling of ultra-thin steel sheet in two-stage tensile deformation considering strain path change and grain size effect and application in multi-stage microforming,” International Journal of Machine Tools and Manufacture, vol. 164, p. 103713, 2021, doi: 10.1016/j.ijmachtools.2021.103713.

[31] X. Wang et al., “Size effects on flow stress behavior during electrically-assisted micro-tension in a magnesium alloy AZ31,” Materials Science and Engineering: A, vol. 659, pp. 215–224, 2016, doi: 10.1016/j.msea.2016.02.064.

[32] H. N. Lu, D. B. Wei, Z. Y. Jiang, D. Wu, and X. M. Zhao, “Study on the influence of temperature on the surface asperity in micro cross wedge rolling,” AIP Conference Proceedings, pp. 1032–1037, 2013, doi: 10.1063/1.4806948.

[33] A. Kumar Pandey and P. P. Date, “Room temperature and high temperature micro-forming analysis of SS304 foil,” Materials and Manufacturing Processes, vol. 37, no. 14, pp. 1691–1700, 2022, doi: 10.1080/10426914.2022.2039697.

[34] M. M. Koochaksaraei, I. Ahmadi, R. Hajian, and M. M. Mohammadi, “Temperature and pressure effects on microchannels dimensions in hot embossing,” Journal of Micromechanics and Microengineering, vol. 32, no. 7, p. 075006, 2022, doi: 10.1088/1361-6439/ac6ec4.

[35] S. Nanthakumar, D. Rajenthirakumar, and S. Avinashkumar, “Influence of temperature on deformation behavior of copper during microextrusion process,” Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, vol. 234, no. 9, pp. 1797–1808, 2020, doi: 10.1177/0954406219899114.

[36] Z. Ren and X. Y. Huang, “Effects of Gases Temperature on the Viscosity and Flow Velocity of Melt in the Gas-Assisted Extrusion Forming of Plastic Micro-Tube,” Materials Science Forum, vol. 976, pp. 145–150, 2020, doi: 10.4028/www.scientific.net/MSF.976.145

[37] C. Wang et al., “Temperature dependent micromechanics-based friction model for cold stamping processes,” Journal of Physics: Conference Series, vol. 1063, p. 012136, 2018, doi: 10.1088/1742-6596/1063/1/012136.

[38] C. Yan, K. Z. Chen, and F. Gong, “Precision Micromassive Forming of Microgears at Different Temperatures,” Strength of Materials, vol. 53, no. 1, pp. 55–64, 2021, doi: 10.1007/s11223-021-00260-8.

[39] Z. Jiang et al., “Influences of temperature and grain size on the material deformability in microforming process,” International Journal of Material Forming, vol. 10, no. 5, pp. 753–764, 2016, doi: 10.1007/s12289-016-1317-4.

[40] Y. Y. Qi et al., “Effects of annealing temperature on micro deep drawing of stainless steel-copper composite,” IOP Conference Series: Materials Science and Engineering, vol. 1270, p. 012092, 2022, doi: 10.1088/1757-899X/1270/1/012092.

[41] B. Eichenhüller and U. Engel, “Microforming of titanium – forming behaviour at elevated temperature,” 4M 2006 - Second International Conference on Multi-Material Micro Manufacture, pp. 339–342, 2006, doi: 10.1016/B978-008045263-0/50077-5.

[42] M. Arentoft, S. Bruschi, A. Ghiotti, N. A. Paldan, and J. V. Holstein, “Microforming of Lightweight Metals in Warm Conditions,” International Journal of Material Forming, vol. 1, no. S1, pp. 435–438, 2008, doi: 10.1007/s12289-008-0088-y.

[43] C. C. Chang and H. S. Chen, “Effect of Grain Size on Micro Deep Drawing of SUS304 Stainless Steel Square Cup,” Key Engineering Materials, vol. 661, pp. 77–82, 2015, doi: 10.4028/www.scientific.net/KEM.661.77
.
[44] Z. Jiang, J. Zhao, and H. Xie, “Simulation of Micro Deep Drawing,” Microforming Technology, pp. 215–239, 2017, doi: 10.1016/B978-0-12-811212-0.00010-8.

[45] G. Wang, Y. Li, S. Liu, J. Yang, and M. Yang, “Micro deep drawing of T2 copper foil using proportional decreased tools,” The International Journal of Advanced Manufacturing Technology, vol. 95, no. 1–4, pp. 277–285, 2017, doi: 10.1007/s00170-017-1111-6.

[46] L. Luo et al., “An experimental and numerical study of micro deep drawing of SUS304 circular cups,” Manufacturing Review, vol. 2, p. 27, 2015, doi: 10.1051/mfreview/2015029.

[47] Z. H. Mahmood, I. K. Irthiea, and K. A. Abed, “Experimental and Simulation investigations of Micro Flexible Deep Drawing Using Floating Ring Technique,” Al-Khwarizmi Engineering Journal, vol. 14, no. 3, pp. 20–31, 2018, doi: 10.22153/kej.2018.12.007.

[48] D. Pan et al., “Effects of forming velocity on micro deep drawing performance with different blank thickness,” Research Square Preprint, 2021, doi: 10.21203/rs.3.rs-838166/v1.

[49] I. K. Irthiea and G. Green, “Evaluation of micro deep drawing technique using soft die-simulation and experiments,” The International Journal of Advanced Manufacturing Technology, vol. 89, no. 5–8, pp. 2363–2374, 2016, doi: 10.1007/s00170-016-9167-2.

[50] F. Jia, J. Zhao, L. Luo, H. Xie, and Z. Jiang, “Experimental and numerical study on micro deep drawing with aluminium-copper composite material,” Procedia Engineering, vol. 207, pp. 1051–1056, 2017, doi: 10.1016/j.proeng.2017.10.1129.

[51] T. Iwama, “Effect of Mechanical Properties and Forming Conditions on Outer Panel Performances of High Strength Steel Sheets,” SAE Technical Paper Series, 2016, doi: 10.4271/2016-01-0355.

[52] Z. Marciniak, K. Kuczyński, and T. Pokora, “Influence of the plastic properties of a material on the forming limit diagram for sheet metal in tension,” International Journal of Mechanical Sciences, vol. 15, no. 10, pp. 789–800, 1973, doi: 10.1016/0020-7403(73)90068-4.

[53] S. K. Singh et al., “Studies on texture and formability of Zircaloy-4 produced by pilgering route,” Journal of Materials Research and Technology, vol. 8, no. 2, pp. 2120–2129, 2019, doi: 10.1016/j.jmrt.2018.11.018.

[54] L. Zhang and J. Wang, “Modeling the localized necking in anisotropic sheet metals,” International Journal of Plasticity, vol. 39, pp. 103–118, 2012, doi: 10.1016/j.ijplas.2012.05.005.

[55] J. Min, T. B. Stoughton, J. E. Carsley, and J. Lin, “Compensation for process-dependent effects in the determination of localized necking limits,” International Journal of Mechanical Sciences, vol. 117, pp. 115–134, 2016, doi: 10.1016/j.ijmecsci.2016.08.008.

[56] D. Priadi, K. Gandjar, and A. Mahmudah, “Research clustering and the state-of-the-art in micro sheet metal forming: a review,” International Journal of Manufacturing Research, vol. 14, no. 4, p. 1, 2019, doi: 10.1504/IJMR.2019.10019047.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 375-388. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3015

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

AN EVALUATION OF DIRECT AND INDIRECT STRATEGIES FOR HIERARCHICAL NET-LOAD FORECASTING

Oyebola Olasupo*[1], Temitope Raphael Ayodele[2], Ayodeji Samson Olatunji Ogunjuyigbe[3]

[1, 2, 3]Department of Electrical and Electronic Engineering, University of Ibadan, Nigeria.

[1]https://orcid.org/0000-0003-0753-5721, [2]https://orcid.org/ 0000-0001-7789-934X, [3]https://orcid.org/0000-0002-0387-7380

Email: *oolasupo8962@stu.ui.edu.ng, tayodele2001@yahoo.com, s.ogunjuyigbe@ui.edu.ng

ARTICLE INFO

Article History
Received: November 27, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Netload,
Forecasting,
Reconciliation,
Solar,
Time Series.

ABSTRACT

Power systems require accurate and coherent net-load forecasts across all levels of the grid hierarchy. A key methodological choice is between a direct strategy (forecasting net-load holistically) and an indirect strategy (forecasting gross load and renewable generation separately). This paper provides the first systematic evaluation of these two approaches within a hierarchical net-load forecasting framework, including the subsequent impact of statistical forecast reconciliation. Using load data from a three-level electricity network and simulated solar PV generation, this study compares the performance of both strategies using neural network auto-regression forecasting model and five reconciliation techniques. The findings show that the direct forecasting strategy yields statistically significant and more accurate forecasts across the hierarchy. However, the direct approach higher accuracy is coupled with over-forecasting bias. The analysis of reconciliation techniques revealed that while optimal methods improved the direct strategy's accuracy, they failed to correct its bias. Conversely, the simple Bottom-Up method, when applied to the indirect strategy, provided the least biased, though less accurate, overall forecast. The study concludes that while the direct approach is more robust for maximizing forecast accuracy, the final choice of strategy and reconciliation method must be aligned with specific operational objectives.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Accurate forecasting of future electricity demand is crucial for efficient generation, transmission and distribution [1-3]. Net-load is defined as the electricity demand minus renewable energy sources (RES) generation [4]. The dual uncertainties from demand and supply variability complicate net-load forecasting (NLF) [1], making it a challenge in the transition to decarbonised power grids [5-7]. To improve the accuracy of NLF, studies have explored various methodological approaches. These include direct NLF (forecasting net-load holistically) and indirect NLF (forecasting gross load and RES generation separately, then adding to form NLF) strategies [6], [8], application of traditional statistical models, Machine Learning (ML) and Deep Learning (DL) architectures [9], [10], and the use of pre-processing techniques like signal decomposition and feature engineering to handle the non-linearity inherent in net-load data [11], [12]. The integration of RES such as solar PV into power grids across the globe necessitates accurate net-load forecasts at various granularities, from individual feeders and substations to the total system level, for effective planning, scheduling and real-time operations.

The existing body of research has extensively addressed the challenge of improving point net-load forecast accuracy at a single level of aggregation. Direct NLF, which forecasts the net load as a single time series, is widely adopted due to its simplicity and applicability in scenarios with limited observability of behind-the-meter (BTM) RES. In contrast, indirect (or additive) forecasting involves separately predicting load and RES generation before computing the net load. Existing literature on direct versus indirect strategies in net load forecasting reports varied outcomes, often due to evaluations limited to a single aggregation level or specific contexts such as microgrids or high renewable penetration systems. For example, Alipour *et al.* [11] integrated deep neural networks with wavelet transforms and found the indirect strategy more accurate, with improvements in metrics like mean absolute percentage error when using transformed historical inputs.

[13] applied LSTM models in microgrids with solar and wind, reporting lower MAE, MSE, RMSE, and NRMSE for the indirect method compared to direct. [14] proposed a disaggregation based additive framework using dictionary learning and LSTM, achieving RMSE reductions of 9.41% to 16.01% and MAE reductions of 9.34% to 15.24% over direct forecasts in neighbourhoods with high BTM PV and low observability. [15] used unsupervised contrastive disaggregation followed by models like DeepAR and TFT, demonstrating superior accuracy for the additive approach over direct net load forecasting. [16] tested machine learning models with and without BTM PV data, noting slight advantages for indirect but minimal differences overall. [17] compared additive (separate load and solar) versus integrated (solar as input to net load model) approaches, finding the integrated method superior due to cointegration of errors. Conversely, some studies support direct strategies.

[18] employed BNNs in renewable microgrids and reported lower NRMSE (3.98%) for direct compared to indirect (4.17%). [19] compared aggregation levels, identifying partially aggregated (separate total demand and supply) as optimal over fully direct or disaggregated. These variations stem from differences in models, data granularity, renewable types, and metrics, with limited attention to hierarchical structures or bias alongside accuracy. While direct and indirect NLF debate is well-documented in the literature, are well-investigated, their application to hierarchical time series, which naturally represent multi-level power systems, remains unexplored in NLF literature. When forecasts are generated independently for each level of the grid, the resulting set of forecasts is almost certainly incoherent which means that the sum of the forecasts for lower-level nodes (e.g., feeders) will not be equal to the forecast for their parent node (e.g., a substation). This incoherence creates significant operational challenges.

The transmission system operator, relying on an aggregate forecast, may make dispatch decisions that conflict with the needs identified by a distribution operator, who is relying on substation-level forecasts. Incoherent forecasts also lead to suboptimal scheduling of ancillary services, energy storage and demand response, as the perceived need at the aggregate level does not align with the sum of needs at the local level. Forecast reconciliation techniques ensure this coherence by adjusting independent base forecasts [20]. The primary contribution of this paper is an evaluation of direct and indirect strategies for hierarchical net-load forecasting (HNLF) and reconciliation. Unlike prior studies which are often limited to a single model or aggregation level, this work compares these two approaches for a hierarchical electricity network with solar photo-voltaic (PV) penetration, using three forecasting models and five statistical reconciliation techniques. The findings of this study is important for identifying effective forecasting strategies in RES-dominated power systems.

III. MATERIALS AND METHODS

I. Data: The primary datasets used in this study are the gross electricity load data and exogenous data. The gross load data for this study was obtained from an electricity network which covers a significant part of a state in the south-western part of Nigeria. Nine distribution feeders at 33kV voltage level are energised via the secondary side of three 132/33kV transformers at a 132/33kV substation. The hierarchical structure of the network is as shown in Figure 1. Historical hourly electricity load data for the feeders are acquired between 00:00 hours of 1st January 2022 to 24:00 hours of 31st December 2024. The raw data is validated and cleaned, while missing values and outliers are addressed. After cleaning, let $\mathbf{y}_t = (y_{1,t}, y_{2,t}, \dots, y_{m,t})^\top \in \mathbb{R}^m, t = 1,2, \dots, T$ be the vector of hourly electricity load observations at the m bottom-level nodes (e.g., feeders) over a horizon of T hours. Then, the complete multi-level load vector is $\mathbf{Y}_t = \mathbf{S}\mathbf{y}_t, \mathbf{Y}_t \in \mathbb{R}^n$, where $n \geq m$ and $\mathbf{S} \in \{0,1\}^{n \times m}$ denote the summing matrix that maps the m bottom-level series to n upper level series (such as transformers and substations).

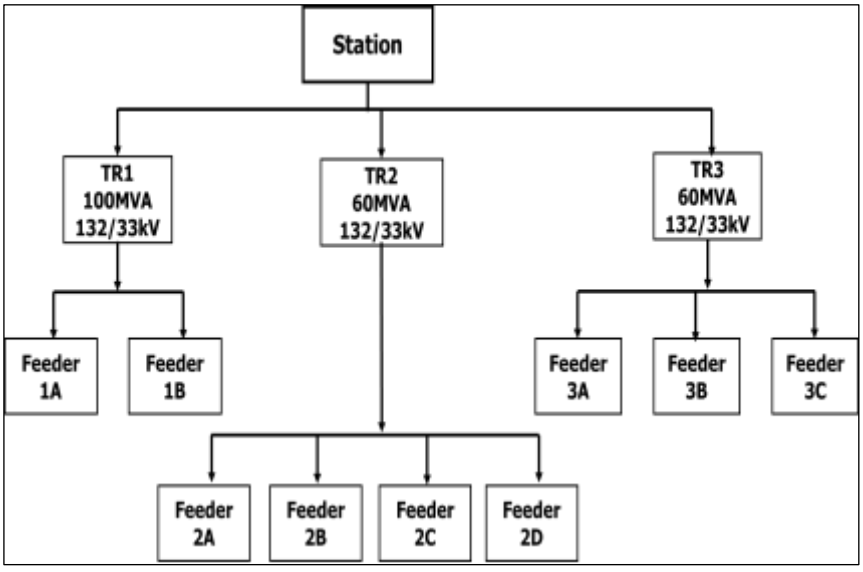


Figure1: Hierarchical Structure of the Electricity Network.
Source: Authors, (2026).

Exogenous data in this study encompasses external variables that influence net load calculations. The exogenous data include hourly meteorological and temporal datasets for the specific geographical location covering the entire historical period, $\mathbf{x}_t = (x_{1,t}, x_{2,t}, \dots, x_{p,t})^\top \in \mathbb{R}^p$). The Global Horizontal Irradiance (GHI, measured in W/m²), was obtained from the NASA POWER database (https://power.larc.nasa.gov/) , providing hourly values that quantify the solar radiation incident on a horizontal surface.

These data were integrated with local meteorological records which was obtained from *Visual Crossing* weather service (https://www.visualcrossing.com/) and include ambient air temperature (°C), precipitation type, weather conditions (Table1). Missing values in the combined dataset were imputed using linear interpolation for numeric variables, ensuring a complete time series without gaps. Categorical meteorological variables (e.g., precipitation type, conditions) were transformed into binary indicators via one-hot encoding [21], creating separate features for each unique category across the dataset. Features with near-zero variance were excluded to reduce redundancy, and all exogenous variables were normalized to a [0, 1] range using min-max scaling [22], based on training period to prevent data leakage. Table 1 summarises the key meteorological variables, their descriptions, types and unit.

Table 1: Meteorological Features.

Feature Name	Description	Type	Unit
GHI_W_m2	Global Horizontal Irradiance	Numeric	W/m²
temp	Ambient air temperature	Numeric	°C
preciptype	Type of precipitation (e.g., rain, none).	Categorical	N/A
conditions	Cloud conditions (e.g., clear, cloudy).	Categorical	N/A
icon	Weather icon descriptor (e.g., sunny, rainy)	Categorical	N/A
stations	Reporting weather station identifiers	Categorical	N/A
name	Location or station name	Categorical	N/A
dew	Dew point temperature	Numeric	°C.
humidity	Relative humidity	Numeric	%
precip	Precipitation amount	Numeric	mm
precipprob	Probability of precipitation	Numeric	%
windgust	Maximum wind gust speed	Numeric	km/h
windspeed	Mean wind speed	Numeric	km/h
winddir	Wind direction	Numeric	degrees
cloudcover	Cloud cover percentage	Numeric	%
solarradiation	Solar radiation	Numeric	W/m²
solarenergy	Solar energy accumulation	Numeric	MJ/m²
uvindex	Ultraviolet radiation index.	Numeric	N/A

Source: Authors, (2026).

To capture temporal and seasonal patterns inherent in load and PV dynamics, additional analytical features were engineered. These include basic components such as year, month number, day of the week, hour of the day, day of the year, and week of the year. Cyclic patterns were encoded using trigonometric transformations to represent periodicity continuously, avoiding discontinuities in numerical representations (Table 2). These exogenous features form the regressors for modelling, enabling the incorporation of weather-driven variability and temporal structures into net load simulations.

Table 2: Temporal Features.

Feature Name	Description	Type
year	Year	Numeric
month_num	Month of the year	Numeric
day_of_week_num	Day of the week,	Numeric
hour_of_day_num	Hour of the day	Numeric
day_of_year_num	Day of the year	Numeric
week_of_year_num	Week of the year	Numeric
hourofday_sin	Sine transformation for the hour of the day	Numeric
hourofday_cos	Cosine transformation for the hour of the day	Numeric
dayofweek_sin	Sine transformation for the day of the week	Numeric
dayofweek_cos	Cosine transformation for the day of the week	Numeric
month_sin	Sine transformation for the month of the year	Numeric
month_cos	Cosine transformation for the month of the year	Numeric
dayofyear_sin	Sine transformation for the day of the year	Numeric
dayofyear_cos	Cosine transformation for the day of the year	Numeric
period_wet	Binary indicator for the wet season	Binary
is_weekend	Binary indicator for weekends	Binary

Source: Authors, (2026).

II. Solar PV Generation Profile: The hourly GHI, is separated into Direct Normal Irradiance (DNI) and Diffuse Horizontal Irradiance (DHI);

$$\mathrm{GHI}_t = \mathrm{DHI}_t + \mathrm{DNI}_t \cos\left(\theta_{z,t}\right) \tag{1}$$

Where the solar zenith angle $\theta_{z,t}$ is defined in terms of the geographical location latitude ϕ, the solar declination δ_t and the hour angle ω_t.

The direct and diffuse components are then mathematically projected from the horizontal plane onto the defined plane of the PV array. This array has a fixed tilt angle β and azimuth angle γ_p. The total Plane-of-Array (POA) irradiance, $G_{POA,t}$, on the tilted surface is the sum of the three components:

$$G_{\mathrm{POA},t} = G_{b,c,t} + G_{d,c,t} + G_{r,c,t} \tag{2}$$

Where $G_{b,c,t}$ is the direct beam irradiance incident on the collector plane at time t, $G_{d,c,t}$ is the diffuse irradiance on the collector plane and $G_{r,c,t}$ the ground-reflected irradiance on the collector plane. The calculated POA irradiance is converted into a realistic electrical power output. Meanwhile, a PV system is not perfectly efficient due to significant losses during the energy conversion process.

Losses arise from inverter inefficiency in converting DC to AC power, resistive losses in wiring, performance degradation at high temperatures, and energy reduction due to soiling. Therefore, these factors are aggregated into a single de-rating factor called the Performance Ratio (PR). For this model, a PR of 0.85 is used, signifying that 85% of the available light energy is successfully converted to usable AC power.

The power output is;

$$P_t = \frac{G_{\mathrm{POA},t}}{1000} \cdot \mathrm{PR} \cdot A \tag{3}$$

Where A is the panel area in squared meters and the division by 1000 converts W/m^2 to kW/m^2. This equation yields the final PV power generation profile in kilowatts, which serves as the foundational template for the subsequent net load simulation. Any resulting negative values (which can occur from model approximations during non-daylight hours) are set to zero to ensure physical realism. The solar PV generation profile was simulated for a fixed, non-tracking system where the substation is located. The model assumed a south-facing orientation with a shallow panel tilt of 7.38°. The R *solarR* package was used for the implementation [23]. The hourly PV profile serves as a standardised shape for PV generation which will be scaled to achieve a target penetration level at the level of each individual feeder. Let $L_i(t)$ be the hourly gross load on feeder i at hour t, and $P_{PV,\mathrm{raw}}(t)$ be the hourly solar PV system AC output from the standardised annual profile. The renewable penetration level is defined as the total annual energy generated by PV relative to the total annual energy consumed by the gross load for each specific feeder i. The total annual gross load energy for feeder i (denoted $E_{L,i}$) and the total annual energy from the raw PV profile shape (denoted $E_{PV,\mathrm{raw}}$) over the period of interest (one year) are given by:

$$E_{L,i} = \sum_{t=1}^{8760} L_i(t) \tag{4}$$

$$E_{PV,\mathrm{raw}} = \sum_{t=1}^{8760} P_{PV,\mathrm{raw}}(t) \tag{5}$$

The scaling factor λ_i that ensures the total scaled PV energy meets the target P_{target} penetration level for feeder i is given as:

$$\lambda_i = \frac{\left(\frac{P_{\mathrm{target}}}{100}\right) \times E_{L,i}}{E_{PV,\mathrm{raw}}} \tag{6}$$

Applying this feeder-specific scaling factor λ_i to the original hourly PV profile $P_{PV,\mathrm{raw}}(t)$ returns the scaled hourly PV generation $P_{PV,\mathrm{scaled},i}(t)$ for feeder i. This process is repeated for each of the three years (2022-2024). This preserves the diurnal and seasonal shape of the PV generation while adjusting its overall magnitude to meet the feeder's target annual energy contribution. To ensure generation is non-negative:

$$P_{PV,\mathrm{scaled},i}(t) = \max\left(0, \lambda_i \times P_{PV,\mathrm{raw}}(t)\right) \tag{7}$$

Therefore, the net load $P_{\mathrm{Net},i}(t)$ at feeder i with P_{target} energy penetration is given as:

$$P_{\mathrm{Net},i}(t) = L_i(t) - P_{PV,\mathrm{scaled},i}(t) \tag{8}$$

After calculating the net loads for each individual feeder, these feeder-level net loads $P_{\mathrm{Net},i}(t)$ are then aggregated upwards to their respective transformer and ultimately to the station level to form the complete net load hierarchical time series. This process ensures that the PV impact is simulated based on the specific load characteristics at the lowest level before being combined into higher-level net load series.

III. Forecast Model: Artificial neural networks (ANNs) represent a class of forecasting methods inspired by simplified mathematical models of the human brain, enabling the capture of complex nonlinear relationships between response variables and predictors [24]. At their core, ANNs consist of interconnected *neurons* organized into an input layer for predictors, an output layer for forecasts, and optional intermediate hidden layers. Simple ANNs without hidden layers are equivalent to linear regression models, where weights are optimized via learning algorithms to minimize cost functions such as mean squared error (MSE). The addition of hidden layers introduces nonlinearity through activation functions, such as the sigmoid, which combine weighted inputs and enhance robustness to outliers [25]. Weights are typically initialized randomly, constrained by a decay parameter to prevent overfitting, and refined through multiple training iterations, with results averaged for stability.

Hyperparameters, including the number of hidden layers and neurons, are often determined via cross-validation. For time series forecasting, NNAR models incorporate lagged values as inputs, analogous to linear autoregressive models [26], [27]. An NNAR(p,k) denotes p lagged inputs and k hidden neurons, while seasonal variants, NNAR$(p,P,k)_m$, include additional lags from prior seasons, m being the seasonal period of the time series. These models can replicate ARIMA structures without stationarity constraints, allowing for asymmetric cyclic patterns. The NNETAR() function from the *forecast* package is used for the implementation. The function automates lag selection for non-seasonal data or seasonally adjusted series. Forecasting proceeds iteratively, using historical data and prior predictions as inputs. External regressors are used, turning the model into a NNARX (NNAR with exogenous inputs).

IV. Forecast Reconciliation: A Hierarchical Time Series (HTS) is a multivariate time series with linear aggregation constraint [28]. Mathematically, an HTS can be represented as;

$$y_t = Sb_t \tag{9}$$

Where y_t is a vector comprising the *n* time series in all levels of the HTS at time *t*, b_t is a vector of the *m* most disaggregated series of the HTS and *S* is the summing matrix [29]. Given an HTS having a total of *n* time series out of which n_b are the bottom-level series, the reconciled forecast is given as;

$$\tilde{y}_h = SG_h\hat{y}_h \tag{10}$$

Where $\tilde{y}_h$ is the reconciled forecasts, $\hat{y}_h$ is the h-step ahead unreconciled (base) forecasts and G_h is an $n_b \times$ n matrix. The function of the G matrix is to map the unreconciled forecasts into the bottom level which are then summed up by *S* to yield the reconciled forecasts $\tilde{y}_h$. A constraint on *G* to make sure the unbiased base forecasts $\hat{y}$ results in an unbiased reconciled forecasts $\tilde{y}_h$ is given by $SGS = S$ [30]. The challenge of statistical forecast reconciliation is to finding the optimal G matrix. This can be achieved by single-level approach or by optimal forecast reconciliation approach. In the Bottom-Up (BU) approach, the forecast of the lowest level series is extracted by *G* and then aggregated by the summing matrix *S* to obtain forecast of the top-level time series.

For the BU, $G = \left[0_{n_h \times n_a}\ I_{n_h}\right]$, where *n* is the total number of series in the hierarchy, n_b is the number of the bottom-level series and the upper-level series are n_a = *n*-n_b, 0 and I are matrix of zeros and the identity matrix respectively. Building up the HTS from the most disaggregated level in BU method means that no information is lost. However, the bottom level series are usually noisy and poses forecasting accuracy challenges. Let V_h and W_h be the variance-covariance matrix of $\tilde{y}_h$ and $\hat{y}_h$with respect to the actual value $\boldsymbol{y}$ respectively. Then, $V_h = SGW_hG'S'$ and the *G* that minimises the error variance of $\tilde{y}_h$ is the one that minimises the trace of V_h which is found via optimisation to be $G = (S'W_h^{-1}S)^{-1}S'W_h^{-1}$[31]. Obtaining W_h is practically difficult and four simplifying assumptions have been presented in the literature. In the following section, $k_h > 0$;

I. Ordinary Least Squares (OLS) Reconciliation [30]: In this approach, $W_h = k_hI$, hence $G = (S'S)^{-1}S'$. This in effect makes G independent of the data and thus makes OLS reconciliation relatively faster to compute. However, the OLS method gives equal weights to all the constituent time series and does not consider inter-relationships within the HTS.

II. Weighted Least Square (WLS) Reconciliation [32]: Here, $W_h = k_h diag(\widehat{W})$ where $\widehat{W} = \frac{1}{T}\sum_{t=1}^{T} e_t\, e_t'$ and e_t is an n-dimensional vector of forecast residuals arranged in the same sequence as the data.

III. Structural Scaling (STR) Reconciliation [27]: Here, $W_h = k_h$diag$(S1)$where 1 is an *m*-dimensional unit vector. The combination weights therefore reflect the structure of the HTS as only the total source of error variance at each node is considered.

IV. Minimal Trace (MinT) [31]: Here $W_h = k_hW_1\boldsymbol{W_h} = \boldsymbol{k_hW_1}$, where $\boldsymbol{W_1}$ is the full one-step covariance matrix with the assumption that the covariance matrices are proportional to each other. Two ways of calculating $W_1$are by calculating the sample covariance (SAM) or to use a shrinkage estimator (SHR).

The R *FoReco* package was used in the implementation of the five reconciliation methods [33].

V. Addressing Negative Values in the Reconciled Forecasts: Net-load becomes negative once the power generated by the solar photovoltaic (PV) system exceeds the gross load. Since negative net-load is sometimes prohibited in non-battery-based systems (i.e., curtailment), the set-negative-to-zero (*sntz*) heuristic was applied to ensure the generated reconciled forecasts are non-negative [33]. The *sntz* heuristic is a simple post-processing step to ensure non-negative, coherent forecasts in a forecast hierarchy. It starts with reconciled forecasts from any standard method (which may include negatives) and adjusts only the bottom-level forecasts before re-aggregating via bottom-up. Let $\tilde{\mathbf{b}}_h$ be the reconciled bottom-level forecasts (potentially with negatives). The *sntz*-adjusted bottom-level forecasts are $\tilde{\mathbf{b}}_h^+ = \max(\tilde{\mathbf{b}}_h, \mathbf{0})$ and the final non-negative reconciled forecasts are $\tilde{\mathbf{y}}_h^+ = \mathbf{S}\tilde{\mathbf{b}}_h^+$. The method may however introduce slight bias compared to optimization-based methods, but it is computationally faster and often yields similar or better NRMSE in practice.

VI. Experimental Setup: Hierarchical gross Load, solar PV generation and net-load at 30% penetration level will be used in this investigation. The net-load forecasting and reconciliation strategy to be evaluated include,

A. Direct Approach (Fig. 2)

I. Form the historical hierarchical net load hierarchy from the historical hierarchical gross load and solar PV generation data.

II. Train one of the forecast models and generate hierarchical net-load forecast.

III. Apply a reconciliation method to this set of forecasts to enforce hierarchical consistency. The result is the first final forecast set.

B. Indirect Approach (Fig. 3)

I. Train one of the forecast models on the historical hierarchical gross load data and generate hierarchical gross load forecast.

II. Train the same forecast models on the historical hierarchical solar PV generation data and generate hierarchical solar PV generation forecast.

III. Subtract the generated hierarchical solar PV generation from the generated hierarchical gross load forecast to obtain the generated hierarchical net-load forecast

IV.Apply a reconciliation method to the hierarchical net-load forecast to enforce hierarchical consistency. The result is the first final forecast set.

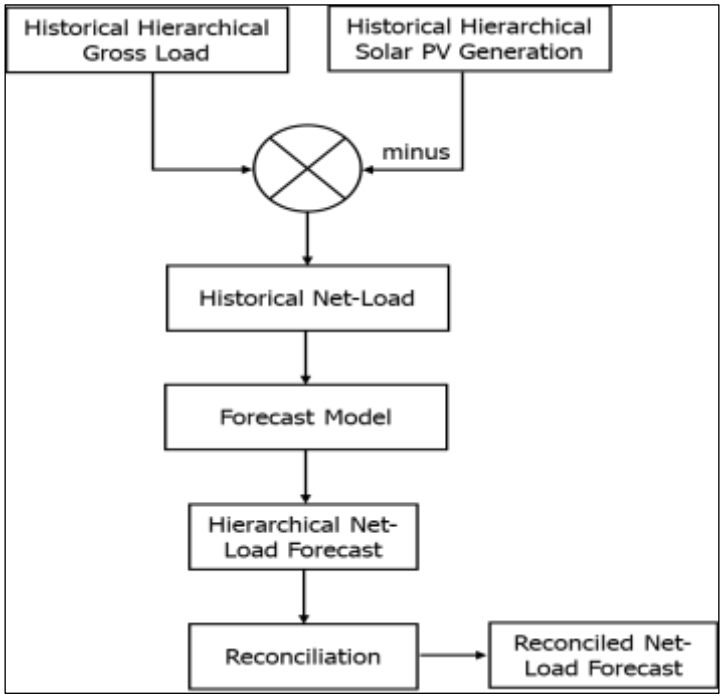


Figure 2: Direct Approach to HNLF.
Source: Authors, (2026).

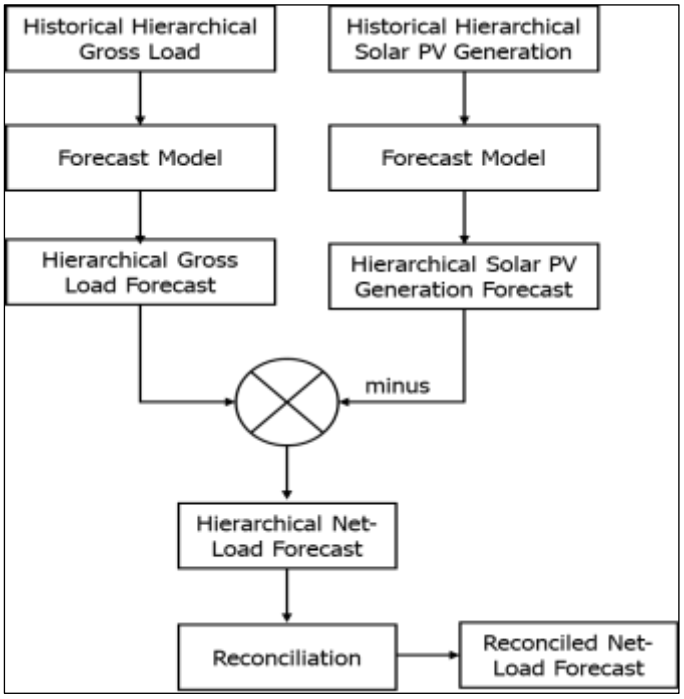


Figure 3: Indirect (Additive) Approach to HNLF.
Source: Authors, (2026).

VII. Evaluation: The final phase is the evaluation which employs a rolling forecast methodology (Fig 4), a technique designed to simulate real-world operational conditions. Instead of using a single, static train-test split, this approach iteratively creates multiple evaluation periods. In each iteration, the models are retrained on a fixed-size window of the most recent historical data (24 months) and then tested on the immediately following time period (one month). This entire process is repeated as the training and testing windows systematically slide forward through time. By aggregating the performance metrics from all iterations, this method provides a reliable assessment of a model's consistency and its true out-of-sample accuracy, ensuring that the results are not dependent on a single, arbitrary evaluation period. The hierarchical nature of the time series is an important factor while selecting metrics for evaluating forecast accuracy. Since series below aggregates to those above, the entire HTS will be on different scales, making absolute metrics not to be appropriate. Thus, this study adopts the Mean Absolute Scaled Error (MASE), Normalised Root Mean Square Error (NRMSE), Normalised Mean Bias Error (NMBE) which are scaled error metrics that enable comparison of the performances of forecasts across different aggregation levels.

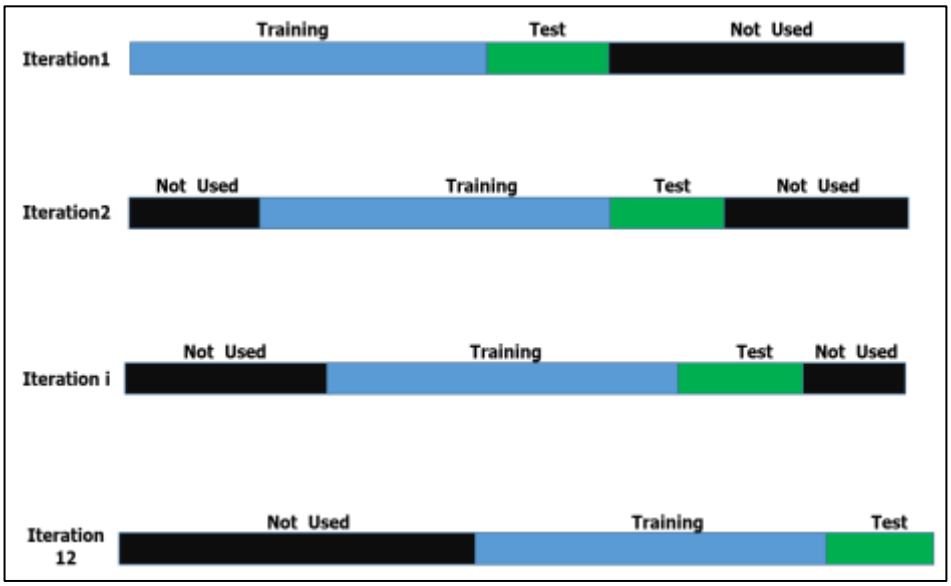


Figure 4: The Rolling Forecast Evaluation Strategy.
Source: Authors (2026).

These metrics have been used in reporting performance of HTS methods [33], [34], [35], [36]. Letting y_t be the actual values, $\hat{y}_t$ the forecasts, m the seasonal period, and n the length of the forecasts;

$$MASE = \frac{\frac{1}{n}\sum_{t=1}^{n} | y_t - \hat{y}_t |}{\frac{1}{T-m}\sum_{t=m+1}^{T} | y_t - y_{t-m} |} \quad (11)$$

Similarly,

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}}{\frac{1}{n}\sum_{t=1}^{n} y_t} \quad (12)$$

Also,

$$NMBE = \frac{\sum_{t=1}^{n} y_t - \hat{y}_t}{n \times \bar{y}} \quad (13)$$

Where T is the length of the in-sample data, $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$ is the mean of the actual series.When evaluating the performance of two forecasting models, statistical tests are essential to determine if one forecast is significantly more accurate than the other [37-39]. These tests typically analyse differences in loss functions, such as mean squared error (MSE) or absolute error, across a set of observations. The Wilcoxon signed-rank test is a non-parametric hypothesis test designed for comparing two related or paired samples to determine if there is a significant difference in their central tendencies [37]. Introduced as an alternative to the parametric paired t-test, it is particularly useful when the data do not satisfy normality assumptions or when the sample size is small. The null hypothesis is that the median of the differences is zero (no systematic difference in forecast accuracy), against the alternative that it is not. To determine if the strategies produced significantly different results, the Wilcoxon signed-rank test was applied. Rather than testing the final aggregated metrics, the test was performed on the paired distributions of the underlying, point-wise error components that constitute the metrics. Specifically, for MASE, the test compared the distributions of absolute errors (|actual - forecast|); for NRMSE, it compared the distributions of squared errors ((actual - forecast)²) and for NMBE, it compared the distributions of raw errors (actual - forecast).

IV. RESULTS AND DISCUSSIONS

I. Data Characteristics: Figure 5 shows the hourly profiles of the three key electrical quantities at the Station level for one full week (January 1st to 7th). The Net Load profile embodies the volatile characteristics of the gross load modulated by the solar PV profile. An indirect approach decomposes the net-load into components at the risk of summing their independent forecast errors, whereas the direct approach models the net-load, avoiding error compounding but tasking the model with learning the entangled component dynamics. The statistical profile of the grid's net load transforms significantly with increasing solar PV penetration (Table 3). While the mean and median net load predictably decrease across all hierarchical levels, this is contrasted by a sharp increase in volatility, as measured by the standard deviation. This heightened variability is a direct consequence of the *duck curve* phenomenon (Fig. 6), where deep midday load suppression creates a greater spread between the low solar-hours demand and the high evening demand [40]. This phenomenon appears to be more pronounced during the dry season than the wet season while being independent of day type (weekend or weekdays).]

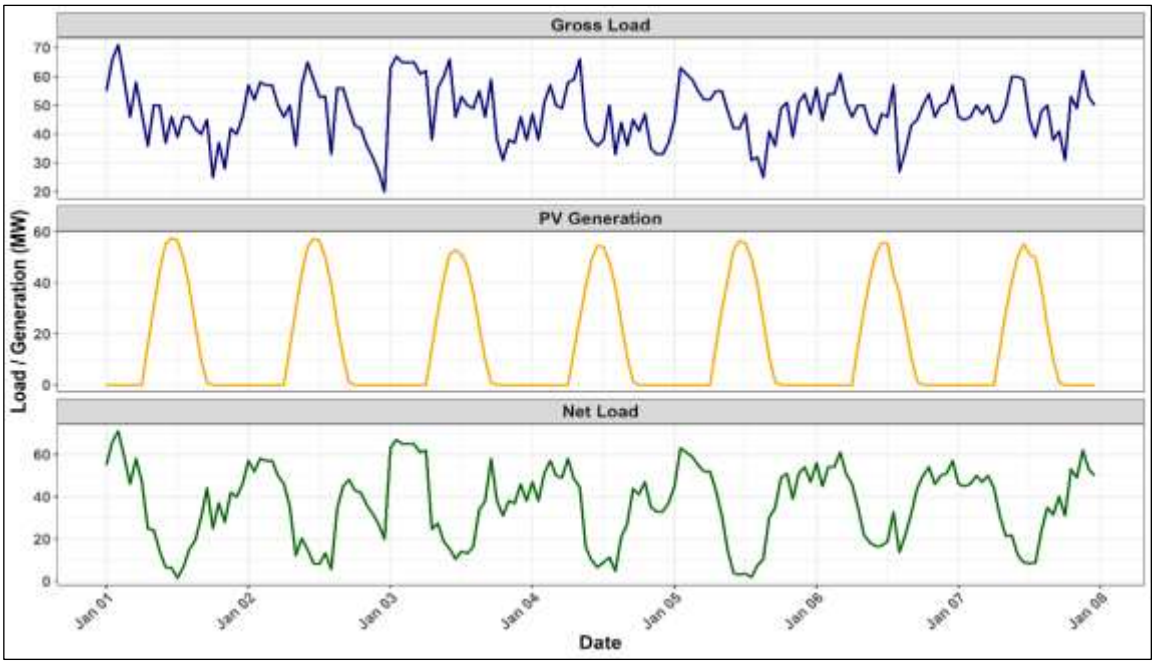


Figure 5: Hourly Profiles of the Three Key Electrical Quantities at The Station Level for A Week.
Source: Authors (2026).

Table 3: Statistical Summary of the Net-Load Data Across the Network Hierarchy.

Level	Penetration (%)	Mean (MW)	Std. Dev (MW)	Median (MW)	Min (MW)	Max (MW)
Level 0	0	46	14.7	47	0	103
	10	42.2	15	42.9	0	95.6
	20	38.6	16.8	39	0	92.3
	30	35.6	18.9	36.8	0	91.9
	40	33.2	20.7	35	0	91.5
	50	31.5	21.9	34	0	91.1
Level 1	0	15.3	6.6	15.7	0	37.7
	10	14.1	6.5	14.3	0	35
	20	12.9	6.9	13	0	34.7
	30	11.9	7.4	11.9	0	34.7
	40	11.1	7.8	11	0	34.7
	50	10.5	8.1	10.6	0	34.7
Level 2	0	5.1	3.2	5.5	0	18.3
	10	4.7	3.1	4.7	0	18.1
	20	4.3	3.1	4.3	0	17.9
	30	4	3.2	3.7	0	17.8
	40	3.7	3.2	3.4	0	17.7
	50	3.5	3.3	3.2	0	17.6

Source: Authors (2026).

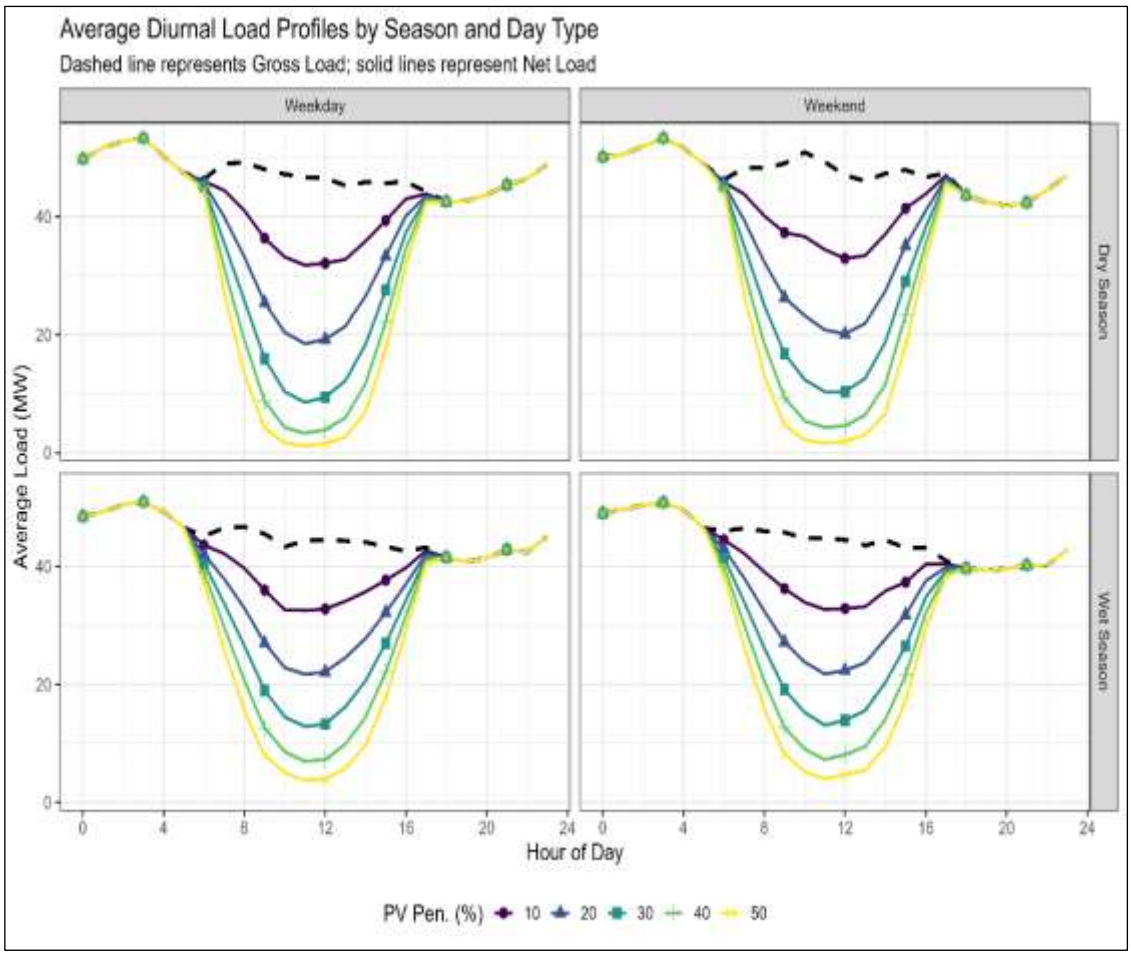


Figure 6: Average Diurnal Net-Load Profiles by Season and Day Type. The Dashed Line Represents the Original Gross Load.
Source: Authors (2026).

PV penetration also alters the predictability of net-load, as evident in the Autocorrelation Function (ACF) plot of Figure 7. The autocorrelation analysis of system net load under varying PV penetrations (0-50%) reveals a dominant 24-hour periodicity consistent with daily load cycles. At 0% PV, the ACF exhibits strong positive correlations at short lags with damped peaks at multiples of 24 hours, reflecting both short-term dependence and daily seasonality. As PV penetration increases, the diurnal cycle in the net load becomes more pronounced, with the ACF showing sharper oscillations and alternating positive (24-hour) and negative (12-hour) correlations.

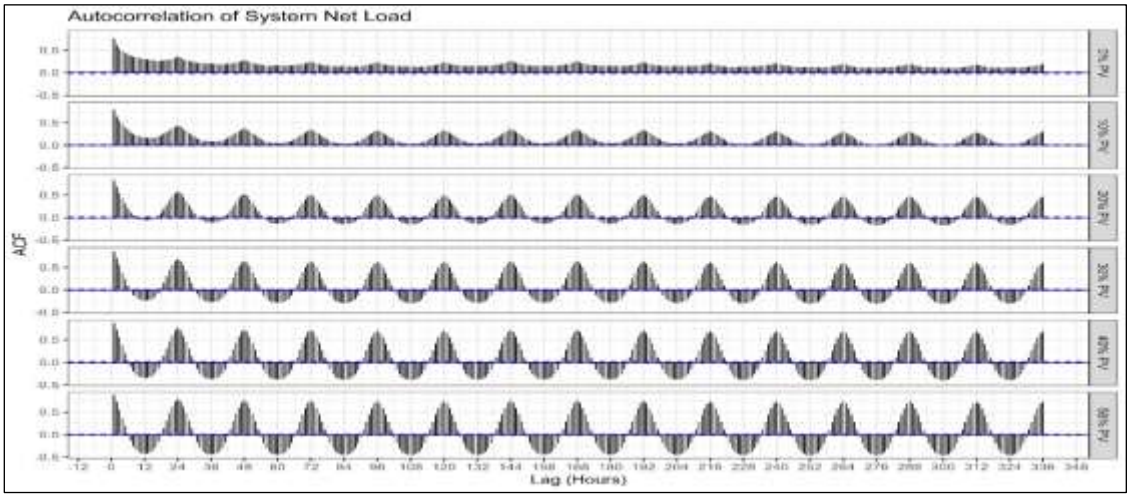


Figure 7: Autocorrelation of The System Net Load. The Correlation at The 24-Hour Lag (And Its Multiples) Is Visibly Weaker for Higher PV Penetration Levels.
Source: Authors(2026).

II. Performance of the Two Strategies: Figure 8 presents a visual analysis of the forecasting model applied to the Station Net Load. It compares a Direct forecasting strategy (Left panel, red colour), where models are trained on the Station Net Load series itself, against an Indirect strategy (Right panel, blue colour), where the Station forecast is derived from other component forecasts. The actual Station Net Load is shown in green for a single representative week. The actual net load exhibits a daily pattern in which the demand is lowest during the early morning hours and peaks in the evening. This cyclical behaviour is the primary feature that both forecasting models are attempting to capture. The Direct forecast successfully captures the overall timing and shape of the daily cycles. It correctly identifies when the peaks and troughs will occur. However, this model exhibits a bias related to the magnitude of the load. At almost every daily peak, the red forecast is noticeably higher than the green actual load. This indicates a tendency to overestimate demand during high-load periods. Conversely, during the lowest points of the day, the forecast is often just higher than the actual load, suggesting it also slightly under-estimates the lows.

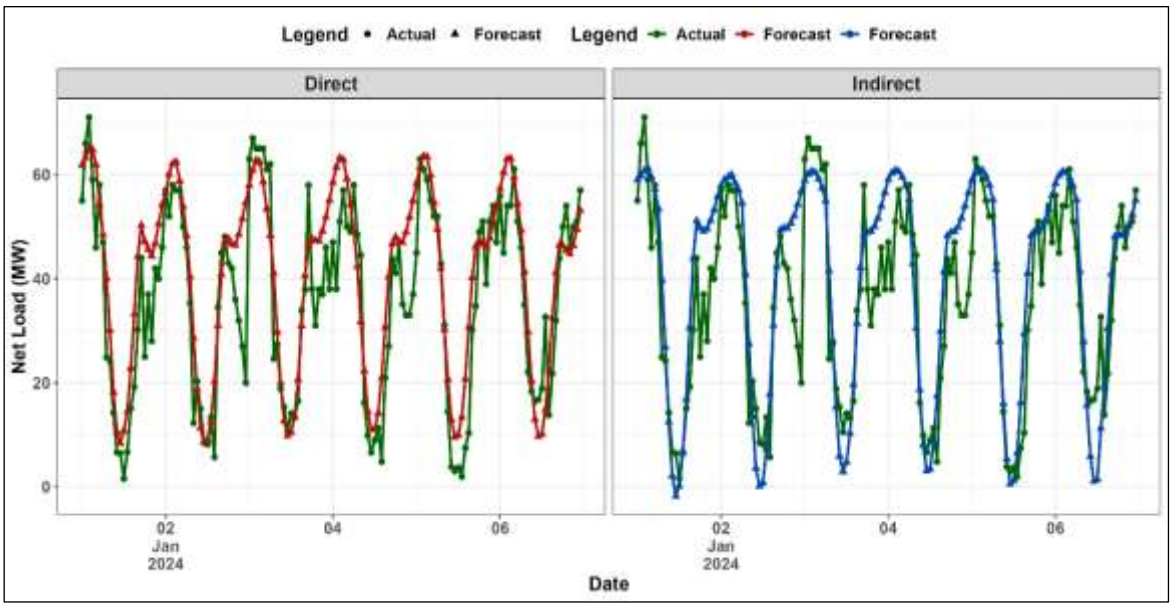


Figure 8: Time Series Plot of the Forecast of the Two Approaches and the Actual Values for the First Week of January 2024.
Source: Authors, (2026).

Like the Direct method, the Indirect forecast also does a good job of tracking the daily rhythm of the net load. However, this model also displays a consistent bias, but in the opposite direction of the Direct method. The Indirect forecast consistently fails to reach the height of the actual peaks. It underestimates demand during high-load periods. Its performance in the troughs appears often lower to the actuals compared to the Direct method. The Indirect model's forecast therefore appears smoother and more conservative. It is less responsive to the sharp peaks in the actual data, leading to its underestimation. The most important finding from this plot is that the two models have opposing systematic biases at the Station level. The Direct model over-forecasts, while the Indirect model under-forecasts. Neither model is consistently accurate in predicting the magnitude of peak load. The better model depends entirely on the business context and the cost associated with forecasting errors. If the cost of under-supplying power (e.g., needing to buy expensive power on the spot market) is very high, the Direct model's tendency to over-forecast might be preferable as a safer option.

If the cost of over-supplying power (e.g., wasted generation, selling excess at a loss) is the primary concern, the Indirect model's conservative, under-forecasting nature might be more desirable. Figure 9 displays boxplots which visualizes the distribution of forecast errors for the Direct and Indirect forecasting strategies. The Y-axis employs a symmetric logarithmic (symlog) scale, a technique necessary for visualizing distributions centred around zero that contain a wide range of values, including large outliers. This scale remains linear near the origin to allow for a clear comparison of the central tendencies, while compressing the tails to effectively display extreme errors without distortion. The Direct strategy exhibits a negative bias, as indicated by its median error being below the zero line. This demonstrates a systematic tendency to over-forecast the net load. The interquartile range (IQR), represented by the height of the box, is relatively compact which suggests that the middle 50% of its forecast errors are clustered within a tighter range, indicating a higher degree of consistency compared to the Indirect method.

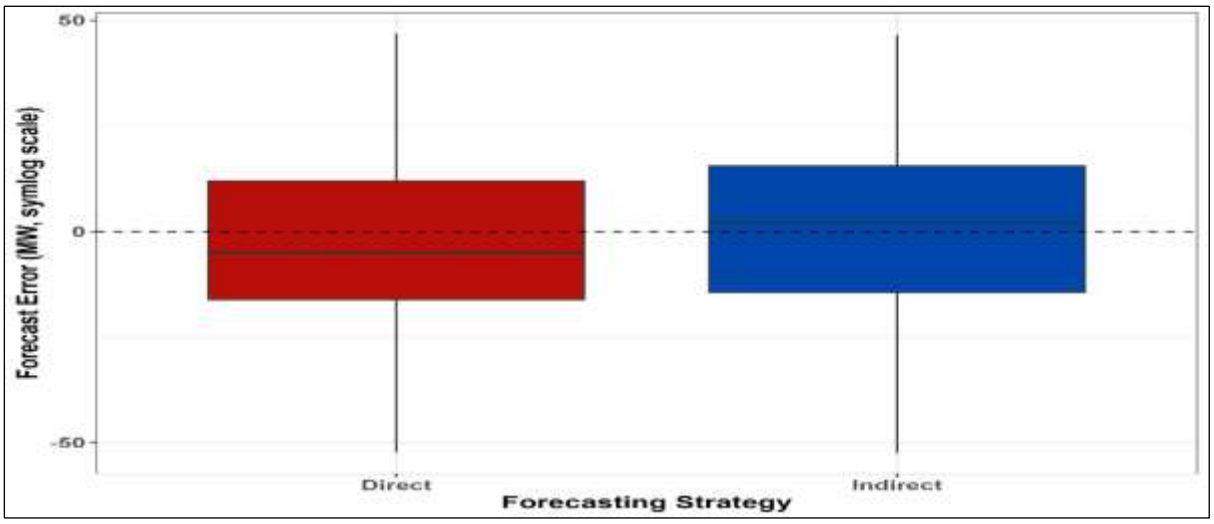


Figure 9: Forecast Error Distribution of the Two Strategies.
Source: Authors (2026).

The whiskers are also symmetrical, showing that the risk of large-magnitude errors is present for both over- and under-forecasting, though the central tendency remains biased towards over-forecasting. The error distribution of the Indirect strategy is more centred, with a median positioned slightly above the zero line. This signifies a minor positive bias, indicating a slight tendency to under-forecast the net load. This strategy displays higher forecast variance as the IQR is visibly larger than that of the Direct method, meaning its forecasts are less consistent, with the central 50% of errors covering a wider range of values. The error distribution is also asymmetrical with the upper whisker notably shorter than the lower one, highlighting a risk of large negative errors. This means the Indirect strategy is more prone to occasional, large-magnitude over-forecasting events. The plot reveals a distinct trade-off between the two forecasting strategies. The Direct method produces more consistent forecasts, but they are subject to systematic over-forecasting bias. In contrast, the Indirect method is less biased on average, but this comes at the cost of higher variance and an asymmetrical risk profile skewed towards major over-forecasting events.

The decision between the two would depend on the specific operational tolerance for predictable bias versus forecast inconsistency and the asymmetrical risk of large errors. The dumbbell plots in Figure 10-12 visualize the impact of different reconciliation methods on the accuracy and bias of the Direct and Indirect base forecasts. Each grey line connects the performance of an unreconciled *Base* forecast (grey point) to its *Reconciled* counterpart (coloured point), immediately revealing the magnitude and direction of the change. An improvement in accuracy is shown by a coloured point to the left of its grey partner, while an improvement in bias is shown by a coloured point closer to the vertical zero line. The MASE plot for both strategies showed that optimal reconciliation methods offer some degree of improvement in terms of accuracy over the base forecast while BU is the worst for both strategies. the Indirect strategy shows the most substantial improvements by the optimal reconciliation methods (Fig. 10). While the SHR method is the overall best for the Direct strategy, the STR method offered the biggest average improvements to the Indirect forecast.

These patterns are sustained in the NRMSE plot (Fig. 11), confirming the validity of the observations. The NMBE plot however tells a completely different story (Fig. 12). While the Indirect method again benefitted more from reconciliation, all reconciled forecast from the Direct strategy remain far from the zero line with reconciliation largely ineffective at correcting the large negative bias (over-forecasting) inherent in this approach. Conversely, for the Indirect strategy, BU is slightly effective at improving bias while optimal reconciliation methods largely worsening the forecast bias. These observations make it clear that the choice of forecasting strategy and reconciliation method must be driven by the primary operational objective. If the goal is maximum forecast accuracy, the plots provide evidence that the Direct strategy combined with SHR or OLS reconciliation is the superior choice. However, if the goal is to produce a less-biased forecast, the Indirect strategy combined with the Bottom-Up (BU) method is the only effective option.

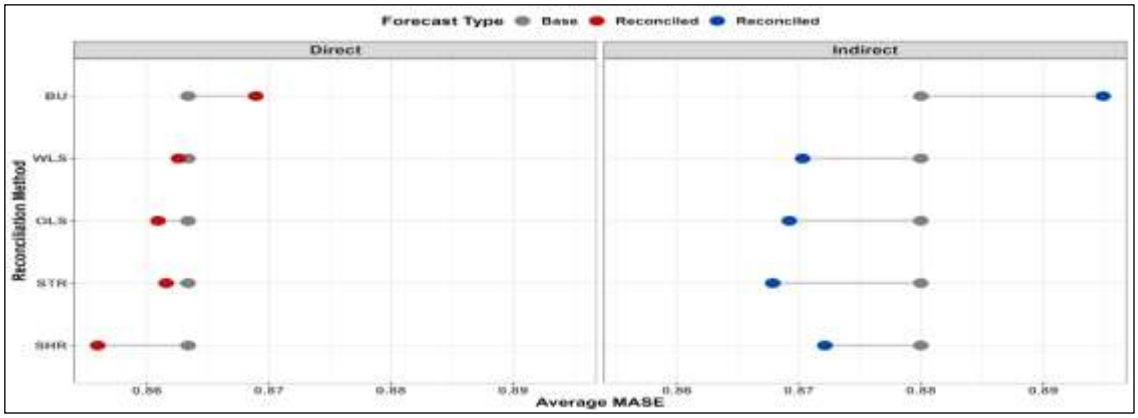


Figure 10: Dumbbell Plots of the MASE of the Direct and Indirect Forecasts.
Source: Authors (2026).

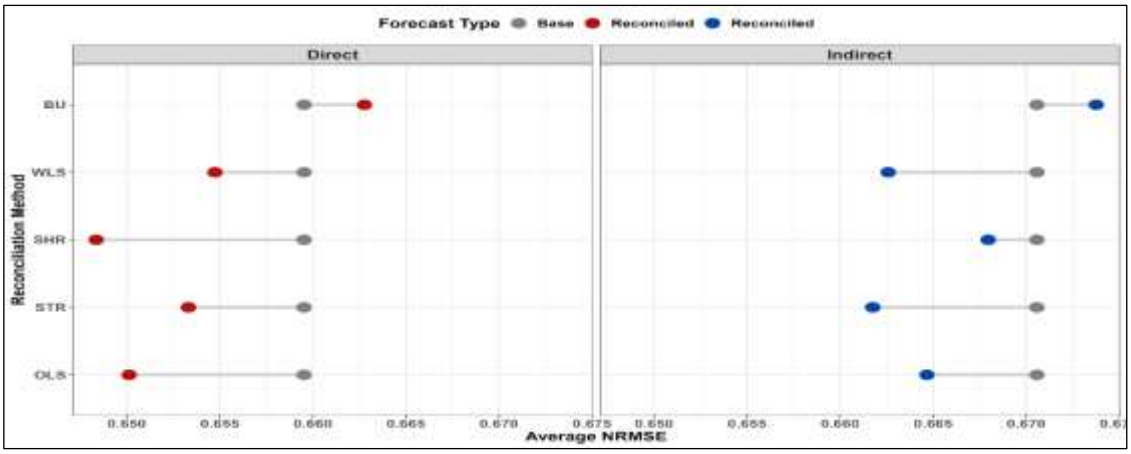


Figure 11: Dumbbell Plots of the NMBE of the Direct and Indirect Forecasts.
Source: Authors (2026).]

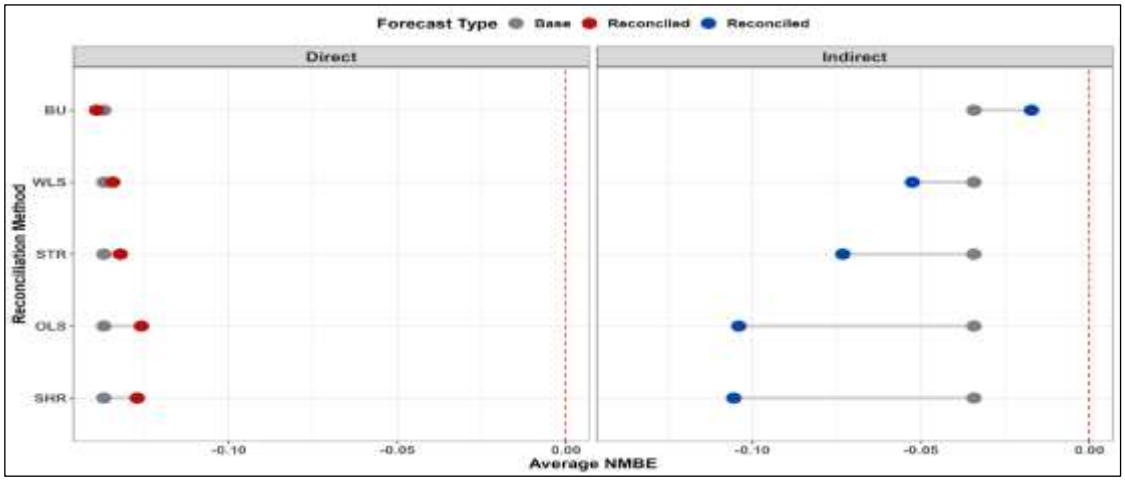


Figure 12: Dumbbell Plots of the NMBE of the Direct and Indirect Forecasts.
Source: Authors (2026).

Table 4 presents the results of a paired Wilcoxon signed-rank test to determine if there is a consistent, significant difference in performance between the Direct and Indirect forecasting strategies. The test is performed independently for each series at each Level of the hierarchy. The columns in the table are interpreted as follows:

I. Series: The series being evaluated.
II. Level: The hierarchical level at which the test was performed.
III. *p*_value: The probability of observing the performance difference between the two strategies if there were no actual underlying difference. A low *p*-value (less than 0.05) suggests that the observed difference is unlikely to be due to random chance.
IV. Median (Direct) / Median (Indirect): The median MASE value for each strategy. The median represents the typical performance, as it is the middle value of all error results and is not skewed by unusually large or small errors. Lower values indicate better typical performance.

V. Significance: A direct interpretation of the *p*-value. If the *p*-value is below the standard threshold of 0.05, the result is declared *Significant*, meaning there is statistical evidence of a real difference between the two strategies. Otherwise, it is *Not Significant*.
VI.Winner: This column identifies the strategy with the better (lower) median MASE, but only if the result is statistically significant. If the difference is not significant, no winner can be declared because the observed difference could be due to random chance.

Table 4: Results of a Paired Wilcoxon Signed-Rank Test to Compare the Two Strategies.

Series	Level	p_value	Median (Direct)	Median (Indirect)	Significance	Winner
Total	Station	1.27E-09	6.906	7.514	Significant	Direct
A	Transformers	1.15E-05	3.447	3.533	Significant	Direct
A1	Feeders	4.40E-87	2.362	2.848	Significant	Direct
A2	Feeders	2.08E-67	2.137	2.439	Significant	Direct
B	Transformers	2.08E-07	3.393	3.377	Significant	Indirect
B1	Feeders	4.13E-65	1.212	1.425	Significant	Direct
B2	Feeders	4.45E-32	0.511	0.599	Significant	Direct
B3	Feeders	1.49E-45	1.991	1.853	Significant	Indirect
B4	Feeders	1.0E-148	1.744	2.178	Significant	Direct
C	Transformers	1.83E-31	3.579	4.181	Significant	Direct
C1	Feeders	9.80E-04	2.020	2.204	Significant	Direct
C2	Feeders	1.40E-82	1.788	1.994	Significant	Direct
C3	Feeders	1.68E-03	1.358	1.456	Significant	Direct

Source: Authors (2026).

The result confirmed that the difference in performance between the two strategies is statistically significant for every single series analysed. All p-values are very small, indicating that the observed differences in accuracy are not due to random chance but reflect a genuine performance gap. With statistical significance established across the board, the *Winner* column clearly identifies the superior strategy based on the lower median absolute error. The Direct strategy is the overall winner, demonstrating superior performance at the highest aggregate level and for the vast majority of the individual components. At the Station level, The Direct method is significantly more accurate, with a median absolute error of 6.91 compared to the Indirect method's 7.51. The Direct strategy wins 2 out of the 3 transformer series (A and C) while the Indirect strategy shows a competitive edge for series B.

The dominance of the Direct strategy is most pronounced at the most disaggregated level. It is the statistically significant winner for 8 out of the 9 feeders. The Indirect method was only superior for the B3 feeder. The statistical analysis provides a definitive conclusion that the Direct forecasting strategy is demonstrably and significantly more accurate than the Indirect strategy for this system**.** Its superior performance is a consistent and statistically validated characteristic observed at the total system level and across almost all of its constituent parts. While the Indirect method may perform better in isolated cases, the evidence overwhelmingly supports the selection of the Direct strategy as the more reliable and accurate base forecasting approach.

V. CONCLUSIONS

The study provides an empirical answer for the direct versus indirect netload forecasting debate with insights from an hierarchical time series point of view, unlike prior studies which were often limited to a single aggregation level and reported conflicting outcomes [6], [11], [41]. This work demonstrates that for the network studied, the Direct approach yields more accurate base forecasts, potentially because a single model can better capture the non-linear interactions between load and solar PV generation that separate additive models may miss and also avoided the compounding of errors from the two component forecasts. The study also quantified the accuracy-bias trade-off inherent in the choice of reconciliation method, providing a practical framework for operational decision-making.

This research empirically demonstrates that while optimal reconciliation methods are designed to improve coherence and minimise error variance [31], they operate by preserving the unbiasedness of the base forecasts and thus cannot fix an already biased model [30]. The practical implication for grid operators is an evidence-based choice. For strategic planning where minimizing error magnitude is paramount, the Direct strategy combined with SHR reconciliation is demonstrably superior. However, for real-time operational scheduling where an unbiased forecast is critical to prevent systemic over-commitment of generation resources, the Indirect strategy with BU reconciliation presents the most viable low-bias alternative. This directly addresses the need to align forecast evaluation metrics with the specific operational decisions being made [42].

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Oyebola Olasupo, Temitope Raphael Ayodele, Ayodeji Samson Olatunji Ogunjuyigbe.
Methodology: Oyebola Olasupo*, Temitope Raphael Ayodele, Ayodeji Samson Olatunji Ogunjuyigbe.
Investigation: Oyebola Olasupo*, Temitope Raphael Ayodele, Ayodeji Samson Olatunji Ogunjuyigbe.
Discussion of results: Oyebola Olasupo*, Temitope Raphael Ayodele, Ayodeji Samson Olatunji Ogunjuyigbe.
Writing – Original Draft: Oyebola Olasupo*, Temitope Raphael Ayodele, Ayodeji Samson Olatunji Ogunjuyigbe.
Writing – Review and Editing: Oyebola Olasupo*, Temitope Raphael Ayodele, Ayodeji Samson Olatunji Ogunjuyigbe.
Resources: Oyebola Olasupo*, Temitope Raphael Ayodele, Ayodeji Samson Olatunji Ogunjuyigbe.
Supervision: Oyebola Olasupo*, Temitope Raphael Ayodele, Ayodeji Samson Olatunji Ogunjuyigbe.
Approval of the final text: Oyebola Olasupo*, Temitope Raphael Ayodele, Ayodeji Samson Olatunji Ogunjuyigbe.

I.REFERENCES

[1] A. Azeem, I. Ismail, S. M. Jameel, and V. R. Harindran, “Electrical Load Forecasting Models for Different Generation Modalities: A Review,” IEEE Access, vol. 9, pp. 142239–142263, 2021, doi: 10.1109/ACCESS.2021.3120731.

[2] A. Perçuku, D. Minkovska, and L. Stoyanova, “Big Data And Time Series Use In Short Term Load Forecasting In Power Transmission System,” Procedia Comput. Sci., vol. 141, pp. 167–174, 2018, doi: 10.1016/j.procs.2018.10.163.

[3] S. I. Kampezidou and S. Grijalva, “Distribution transformers short-term load forecasting models,” in IEEE Power and Energy Society General Meeting, IEEE Computer Society, Nov. 2016. doi: 10.1109/PESGM.2016.7741174.

[4] Y. Chu, H. T. C. Pedro, A. Kaur, J. Kleissl, and C. F. M. Coimbra, “Net load forecasts for solar-integrated operational grid feeders,” Sol. Energy, vol. 158, pp. 236–246, Dec. 2017, doi: 10.1016/j.solener.2017.09.052.

[5] S. E. Razavi, A. Arefi, G. Ledwich, G. Nourbakhsh, D. B. Smith, and M. Minakshi, “From Load to Net Energy Forecasting: Short-Term Residential Forecasting for the Blend of Load and PV Behind the Meter,” IEEE Access, vol. 8, pp. 224343–224353, 2020, doi: 10.1109/ACCESS.2020.3044307.

[6] G. Aburiyana, H. Aly, and T. Little, “Direct Net Load Forecasting Using Adaptive Neuro Fuzzy Inference System,” in 2021 IEEE Electrical Power and Energy Conference (EPEC), Toronto, ON, Canada: IEEE, Oct. 2021, pp. 131–136. doi: 10.1109/EPEC52095.2021.9621457.

[7] X. Sun and C. Jin, “Impacts of Solar Penetration on Short-Term Net Load Forecasting at the Distribution Level,” in 2021 IEEE 4th International Electrical and Energy Conference (CIEEC), Wuhan, China: IEEE, May 2021, pp. 1–6. doi: 10.1109/CIEEC50170.2021.9510770.

[8] S. Kerkau, S. Sepasi, H. O. R. Howlader, and L. Roose, “Day-Ahead Net Load Forecasting for Renewable Integrated Buildings Using XGBoost,” Energies, vol. 18, no. 6, p. 1518, Mar. 2025, doi: 10.3390/en18061518.

[9]S. Sreekumar, K. C. Sharma, and R. Bhakar, “Grey System Theory Based Net Load Forecasting for High Renewable Penetrated Power Systems,” Technol. Econ. Smart Grids Sustain. Energy, vol. 5, no. 1, p. 21, Dec. 2020, doi: 10.1007/s40866-020-00094-4.

[10] P. Kobylinski, M. Wierzbowski, and K. Piotrowski, “High-resolution net load forecasting for micro-neighbourhoods with high penetration of renewable energy sources,” Int. J. Electr. Power Energy Syst., vol. 117, p. 105635, May 2020, doi: 10.1016/j.ijepes.2019.105635.

[11] M. Alipour, J. Aghaei, M. Norouzi, T. Niknam, S. Hashemi, and M. Lehtonen, “A novel electrical net-load forecasting model based on deep neural networks and wavelet transform integration,” Energy, vol. 205, p. 118106, Aug. 2020, doi: 10.1016/j.energy.2020.118106.

[12] O. Rubasinghe et al., “Highly accurate peak and valley prediction short-term net load forecasting approach based on decomposition for power systems with high PV penetration,” Appl. Energy, vol. 333, p. 120641, Mar. 2023, doi: 10.1016/j.apenergy.2023.120641.

[13] J. Silva-Rodriguez, E. Raffoul, and X. Li, “LSTM-Based Net Load Forecasting for Wind and Solar Power-Equipped Microgrids,” in 2024 56th North American Power Symposium (NAPS), El Paso, TX, USA: IEEE, Oct. 2024, pp. 1–6. doi: 10.1109/NAPS61145.2024.10741801.

[14] A. Stratman, T. Hong, M. Yi, and D. Zhao, “Net Load Forecasting with Disaggregated Behind-the-Meter PV Generation,” in 2022 IEEE Industry Applications Society Annual Meeting (IAS), Detroit, MI, USA: IEEE, Oct. 2022, pp. 1–5. doi: 10.1109/IAS54023.2022.9940025.

[15] C. Thepprom, N. Nupairoj, and P. Vateekul, “A Deep Learning Framework for Net Load Forecasting With Unsupervised Behind-the-Meter Disaggregated Data,” IEEE Access, vol. 12, pp. 94958–94971, 2024, doi: 10.1109/ACCESS.2024.3425590.

[16] A. Falces et al., “Short-term net load forecast in distribution networks with PV penetration behind the meter,” Energy Rep., vol. 9, pp. 115–122, May 2023, doi: 10.1016/j.egyr.2022.12.103.

[17] A. Kaur, L. Nonnenmacher, and C. F. M. Coimbra, “Net load forecasting for high renewable energy penetration grids,” Energy, vol. 114, pp. 1073–1084, Nov. 2016, doi: 10.1016/j.energy.2016.08.067.

[18] G. Tziolis et al., “Short-term electric net load forecasting for solar-integrated distribution systems based on Bayesian neural networks and statistical post-processing,” Energy, vol. 271, p. 127018, May 2023, doi: 10.1016/j.energy.2023.127018.

[19] M. Beichter, K. Phipps, M. M. Frysztacki, R. Mikut, V. Hagenmeyer, and N. Ludwig, “Net load forecasting using different aggregation levels,” Energy Inform., vol. 5, no. S1, p. 19, Sep. 2022, doi: 10.1186/s42162-022-00213-8.

[20] N. Sharma, R. Bhakar, and P. Jain, “Optimal reconciliation of hierarchical wind power forecasts of correlated wind farms,” Sustain. Energy Grids Netw., vol. 35, p. 101091, Sep. 2023, doi: 10.1016/j.segan.2023.101091.

[21] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, “Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network,” IEEE Trans. Smart Grid, vol. 10, no. 1, pp. 841–851, Jan. 2019, doi: 10.1109/TSG.2017.2753802.

[22] S. Bhanja and A. Das, “Impact of Data Normalization on Deep Neural Network for Time Series Forecasting,” Jan. 07, 2019, arXiv: arXiv:1812.05519. Accessed: Dec. 02, 2024. [Online]. Available: http://arxiv.org/abs/1812.05519

[23] O. Perpiñán, “solaR : Solar Radiation and Photovoltaic Systems with R,” J. Stat. Softw., vol. 50, no. 9, 2012, doi: 10.18637/jss.v050.i09.

[24] I. N. da Silva, D. Hernane Spatti, R. Andrade Flauzino, L. H. B. Liboni, and S. F. dos Reis Alves, Artificial Neural Networks. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-43162-8.

[25] A. D. Rasamoelina, F. Adjailia, and P. Sincak, “A Review of Activation Function for Artificial Neural Network,” in 2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI), IEEE, Jan. 2020, pp. 281–286. doi: 10.1109/SAMI48414.2020.9108717.

[26] H. Qiao, K. Chalermyanont, and R. Duangsoithong, “Hour-Ahead Power Load Demand Time Series Forecasting Using Four Methods in Three Cases,” in 2019 16th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Pattaya, Thailand: IEEE, Jul. 2019, pp. 593–596. doi: 10.1109/ECTI-CON47248.2019.8955222.

[27] R. J. Hyndman and G. Athanasopoulos, Forecasting: principles and practice. Melbourne: OTexts, 2021.

[28] J. P. Karmy and S. Maldonado, "Hierarchical time series forecasting via Support Vector Regression in the European Travel Retail Industry," Expert Syst. Appl., vol. 137, pp. 59–73, Dec. 2019, doi: 10.1016/j.eswa.2019.06.060.

[29] G. Athanasopoulos, R. J. Hyndman, N. Kourentzes, and A. Panagiotelis, "Forecast reconciliation: A review," Int. J. Forecast., vol. 40, no. 2, pp. 430–456, Apr. 2024, doi: 10.1016/j.ijforecast.2023.10.010.

[30] R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, and H. L. Shang, "Optimal combination forecasts for hierarchical time series," Comput. Stat. Data Anal., vol. 55, no. 9, pp. 2579–2589, Sep. 2011, doi: 10.1016/j.csda.2011.03.006.

[31] S. L. Wickramasuriya, G. Athanasopoulos, and R. J. Hyndman, "Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization," J. Am. Stat. Assoc., vol. 114, no. 526, pp. 804–819, Apr. 2019, doi: 10.1080/01621459.2018.1448825.

[32] R. J. Hyndman, A. J. Lee, and E. Wang, "Fast computation of reconciled forecasts for hierarchical and grouped time series," Comput. Stat. Data Anal., vol. 97, pp. 16–32, May 2016, doi: 10.1016/j.csda.2015.11.007.

[33] T. Di Fonzo and D. Girolimetto, "Spatio-temporal reconciliation of solar forecasts," Sol. Energy, vol. 251, pp. 13–29, Feb. 2023, doi: 10.1016/j.solener.2023.01.003.

[34] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," 2005. [Online]. Available: http://www.forecasters.org/data/m3comp/m3comp.htm

[35] D. Yang, H. Quan, V. R. Disfani, and C. D. Rodríguez-Gallegos, "Reconciling solar forecasts: Temporal hierarchy," Sol. Energy, vol. 158, pp. 332–346, Dec. 2017, doi: 10.1016/j.solener.2017.09.055.

[36] G. M. Yagli, D. Yang, and D. Srinivasan, "Reconciling solar forecasts: Sequential reconciliation," Sol. Energy, vol. 179, pp. 391–397, Feb. 2019, doi: 10.1016/j.solener.2018.12.075.

[37] B. Rosner, R. J. Glynn, and M. T. Lee, "The Wilcoxon Signed Rank Test for Paired Comparisons of Clustered Data," Biometrics, vol. 62, no. 1, pp. 185–192, Mar. 2006, doi: 10.1111/j.1541-0420.2005.00389.x.

[38] G. Dudek, "Short-term load forecasting using Theta method," in E3S Web of Conferences, EDP Sciences, Feb. 2019. doi: 10.1051/e3sconf/20198401004.

[39] Z. Zhang, W.-C. Hong, and J. Li, "Electric Load Forecasting by Hybrid Self-Recurrent Support Vector Regression Model With Variational Mode Decomposition and Improved Cuckoo Search Algorithm," IEEE Access, vol. 8, pp. 14642–14658, 2020, doi: 10.1109/ACCESS.2020.2966712.

[40] I. Calero, C. A. Canizares, K. Bhattacharya, and R. Baldick, "Duck-Curve Mitigation in Power Grids With High Penetration of PV Generation," IEEE Trans. Smart Grid, vol. 13, no. 1, pp. 314–329, Jan. 2022, doi: 10.1109/TSG.2021.3122398.

[41] G. Tziolis, A. Livera, A. Michail, G. Makrides, and G. E. Georghiou, "Direct Against Indirect Short-Term Net Load Forecasting Using Machine Learning Principles for Renewable Microgrids," in 2023 IEEE International Smart Cities Conference (ISC2), Bucharest, Romania: IEEE, Sep. 2023, pp. 1–5. doi: 10.1109/ISC257844.2023.10293666.

[42] R. Hollyman, F. Petropoulos, and M. E. Tipping, "Understanding forecast reconciliation," Eur. J. Oper. Res., vol. 294, no. 1, pp. 149–160, Oct. 2021, doi: 10.1016/j.ejor.2021.01.017.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 389-399. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3019

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

A SELF–EVALUATION MATURITY ASSESSMENT MODEL FOR LEAN SIX SIGMA IMPLEMENTATION IN SMES

Sushil S Mishra *[1] and Ravi Terkar[2]

[1]Mechanical Dept, Mumbai University VIVA Institute of Technology Virar east, Mumbai, India.
[2]Mechanical Dept, NMIMS University, India.

[1]https://orcid.org/0000-0002-2364-0743, [2]https://orcid.org/0000-0002-4851-0208

Email: *sushilmishra@viva-technology.org, Ravi.Terkar@nmims.edu

ARTICLE INFO

Article History
Received: November 28, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
SMEs,
Lean Six Sigma,
Maturity Model,
Process improvement,
Industrial Engineering.

ABSTRACT

This study develops a structured maturity model to assess SMEs' readiness for the implementation of Lean Six Sigma. The model presents a tool that will assist organizations in evaluating their current maturity level before LSS deployment. A comprehensive literature review was performed to find the critical success elements, barriers, and relevant frameworks. These insights were then used to develop a four-level LSS maturity model with a scoring-based assessment instrument. The developed model was refined by expert consultations and subsequently validated using four SME case studies. Findings: The model successfully distinguished the preparedness levels of LSS among the participating SMEs. Scores ranged from Level 1 (Uncertain or Limited) to Level 3 (Understanding or Good). The tool demonstrated strong capability in identifying organizational strengths, gaps, and priorities. Practical implications: It can assist owners or managers by providing an evidence-based mechanism to assess their readiness, benchmark their performance, and plan the implementation of LSS more effectively. It aids structured decision-making and reduces the likelihood of failure at early deployment stages. Originality/value: This work is one of the first LSS maturity models focused on SMEs, along with a quantifiable scoring system, making both academic and practical contributions to improve the assessment of LSS.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Manufacturing enterprises today are under high pressure to improve productivity at lower costs along with high quality, in a competitive global environment [1], [2]. In such a scenario, high-quality output is no longer an option—it is a question of survival. To achieve this, most organizations depend upon continuous improvement methods, which are effectively used in the manufacturing industries [3], and Lean Six Sigma, which combines Lean's principles of waste reduction and Six Sigma's statistical and variation control techniques, has been recognized as a structured strategy that can enhance operating efficiency and competitiveness [4], [5]. However, as far as SMEs are concerned, they face unique challenges when implementing LSS, such as limited technical knowledge, informal organization structures, and limited financial resources compared to major organizations [6-8]. Although SMEs could gain considerably from LSS, Current frameworks are generally developed for large organizations and are too complex or resource-intensive for smaller firms [9-11].

Therefore, most SMEs initiate LSS initiatives based on intuition rather than a systematic understanding of their readiness [12]. While recent studies have considered readiness evaluation as the first step for any LSS deployment process, despite this recognition, the literature suggests that there is currently no valid self-assessment instrument that would organizations to measure their maturity level. Most contributions either focus on selective factors, lack any scoring mechanisms, or remain purely conceptual without any diagnostic instruments. To take up this challenge, the present study proposes a maturity model based on self-evaluation specifically designed for SMEs concerning the implementation of Lean Six Sigma. Guided by the model, SMEs would be able to recognize their strengths, weaknesses, and areas where improvement is needed.

The model will support the SMEs in making well-grounded decisions with regard to the introduction and implementation of LSS by offering a structured diagnostic instrument. It aims at reducing reliance on external consultancy, increasing the success rate of LSS implementations, and also anticipating operational and sociocultural hurdles during transformative change. Thus, this study aims to address the following questions:

1. Under what organizational conditions can SMEs be considered ready to implement Lean Six Sigma?
2. How can the readiness or maturity of SMEs for implementing Lean Six Sigma be systematically evaluated?

This research contributes to the further development of Lean Six Sigma scholarship by proposing a maturity self-assessment model, developing and validating it, and thus offering a practical instrument to support SMEs in planning and executing LSS initiatives with improved effectiveness.

II. THEORETICAL REFERENCE

II.1 AN OVERVIEW OF SMALL AND MEDIUM ENTERPRISES AND LEAN SIX SIGMA.

LSS combines Lean's philosophy of waste elimination along with the statistical rigor of Six Sigma in pursuit of improved operational efficiency, process stability, and product [13], [14]. Despite being widely adopted by large organizations, SMEs face different challenges because of their limited financial resources, informal organizational structures, and dependence on tacit worker knowledge [15]. Amongst other researchers, [7] and [16] emphasized that the successful implementation of LSS within SMEs significantly depends upon leadership support, learning culture, internal competencies, and basic infrastructural resources. However, most of the currently existing implementation frameworks are either overly complex, resource-intensive, or intended for large-scale enterprise use and, therefore, limit their practical applicability to SMEs [9]. Encouraged by this need, recent research efforts have pushed toward recommending simple, structured, and SME-specific maturity assessment models [17], [12].

II.1.1 Obstacles and Difficulties for Small and Medium Enterprises

Although Lean Six Sigma has been effective in enhancing quality and operational performance [18], [19], SMEs frequently experience constraints that limit its successful implementation [20]. Table 1 highlights in summary form the primary barriers identified in the literature.

Table 1: Key Barriers to Lean Six Sigma Adoption in SMEs.

Barriers	Description	Sources
Resource Limitations	Some factors include limited time, budget, and manpower to undertake improvement projects	[18],[21]
Lack of LSS knowledge and skills	Employees have limited familiarity with Lean Six Sigma tools and methods	[7],[11]
Poor organizational culture	Resistance to change and inconsistency in CI practice	[10],[16]
Limited leadership commitment	Support for LSS is still more symbolic than strategic	[2], [22]
Inadequate Infrastructure	Lack of harmonization in processes, data systems, and documentation	[12],[23]

Source: Authors, (2026).

II.1.2 Important Success Elements for Small and Medium Enterprises

The effective application of Lean Six Sigma in SMEs is accomplished by some organizational enablers that support structured improvement, learning, and employee engagement. Literature also identifies leadership commitment, workforce involvement, skills development, tool usage, motivation systems, and supportive working conditions as the key enablers for LSS implementation in organizations [18],[24],[25]. The main Critical Success elements, their applicability to SMEs, and their compatibility with the dimensions of the suggested maturity model are shown in Table 2.

Table 2: Essential Components of LSS Success in SMEs.

Essential Components	Relevance in SMEs	Corresponding Maturity Model Dimension	Sources
Top management commitment	Ensures strategic support and resource allocation	Top Management & Employee Involvement	[18], [22]
Employee involvement and empowerment	Encourages participation in CI and project activities	Top Management & Employee Involvement	[7], [25]
Structured training and skill development	Builds capability to apply LSS tools effectively	Infrastructure & Training	[10], [11]
Availability and use of improvement tools	Enables systematic problem-solving and process control	Tools & Techniques	[12], [17]
Reward and recognition mechanisms	Maintains motivation and supports cultural acceptance of CI	Motivation & Recognition	[4],[18]
Fair and safe working environment	Promotes workforce stability and supports an improvement culture	Social Performance	[7],[25],[26]

Source: Authors, (2026).

II.1.3 Foundation for integrating Lean, Six Sigma, and Lean Six Sigma

Significant efforts have been made to develop frameworks and models that help apply Lean and Six Sigma principles in practice. Table 3 highlights a thorough examination of these contributions.

Table 3: List of different lean and Six Sigma models and frameworks.

Sr.no	Authors	Strengths	Limitations	Scoring System
1.	[22]	Emphasizes how crucial it is to do a readiness assessment before implementing LSS. Suggests a paradigm that combines knowledge and risk management.	Early-stage idea; framework requires additional real-world case testing.	No
2.	[11]	Offers a thorough scholarly summary of the LSS deployment variables based on a systematic review.	Theoretical, with little validation from practitioners.	No
3.	[12]	Proposes a streamlined framework for SMEs to integrate Six Sigma and Lean tools.	Validation is restricted to one case study; generalization is problematic.	No
4.	[9]	Identifies the obstacles and readiness criteria for LSS adoption in the context of SMEs.	Lacks a measurement tool that has been validated; the framework is still conceptual.	No
5.	[27]	Highlights important leadership and cultural elements while concentrating on LSS implementation in the service industry.	Limited case validation; less established manufacturing applicability.	No
6.	[28]	Highlights the obstacles and elements that make SMEs successful while implementing LSS. offers helpful advice to managers.	Mostly exploratory, with a greater emphasis on concept than prescription.	No
7.	[29]	Suggests an LSS deployment paradigm that is centered on SMEs and places a strong emphasis on flexibility and ease of usage.	The scope is restricted to Swedish SMEs and requires more extensive empirical validation.	No

Source: Authors, (2026).

An overview of current LSS implementation and readiness models indicates that, while there have been numerous studies that have made useful contributions, the most significant shortcoming is the lack of a standardized scoring system. [22] recognize the importance of readiness assessment and provide a paradigm that includes knowledge and risk management, but that remains at a conceptual level without measurable scoring. Similarly, [11] provide a systematic review of LSS deployment factors but avoid providing a measurement instrument at the practitioner level. Although models by [12], [9], [27] offer useful insights in highlighting challenges, cultural concerns, and SME-specific issues, their frameworks are mostly descriptive in nature and lack validated measurements. Even [17], [29] studies, as informative as they were in pinpointing success factors and barriers, are exploratory and contextual and thus possess limited generalizability. This is the critical gap that highlights the originality of this proposed model in this research. By converting qualitative information into measurable metrics, the scoring-based model improves objectivity, comparability, and usability for practitioners and SMEs.

III. RESEARCH METHODOLOGY AND DEVELOPMENT OF MODEL

Research shows that although Lean Six Sigma research has advanced significantly in industrialized nations [30-32], developing countries are still lacking context-specific adoption frameworks [33]. To address this gap, a practical and SME-friendly self-evaluation maturity model was created over the course of nine months using a systematic four-stage process in order to close this gap (Figure 1).

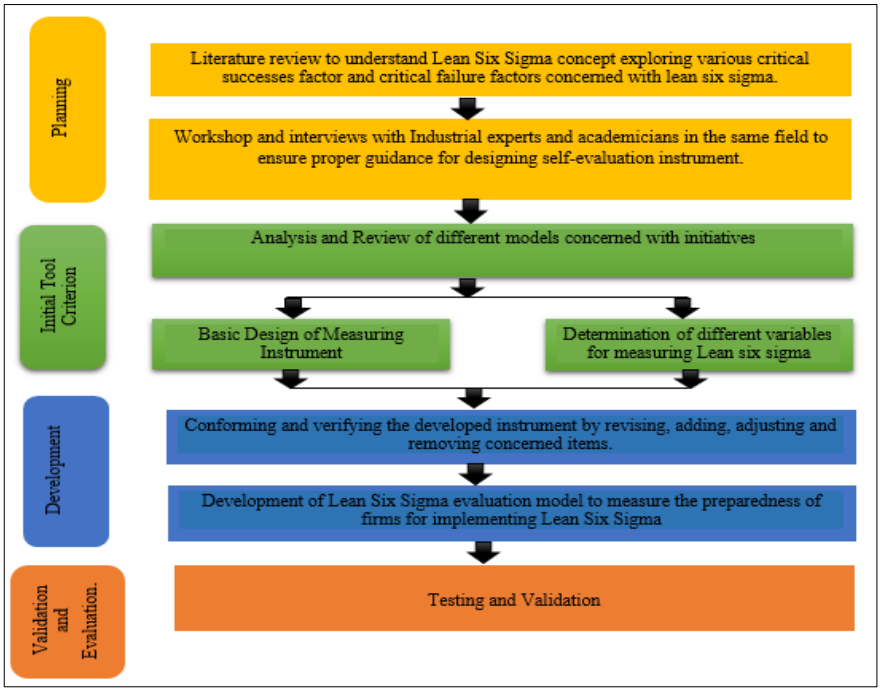


Figure 1: Research methodology.
Source: Adapted from [34].

Planning: In the beginning, five industry professionals participated in semi-structured interviews. All of the professionals who took part had carried out Lean six sigma initiatives within their firms. These professionals all had a cumulative experience in continuous improvement of between 2 and 20 years. A workshop on Lean and Six Sigma was also held, this time with three professors. Through networking and personal connections, industry professionals and academics were located. Although the workshop and interview results aren't discussed in great detail in this section, they provided a strong foundation for the preparation self-evaluation model instrument. The results revealed that there is a vital need to create a realistic instrument to assess whether businesses were ready to implement Lean and Six Sigma practices, because during the study, it was found that there was still a lot of difference between the current industrial scenario and theory related to the evaluation of Lean and Six Sigma implementation. The majority of participants mentioned challenges when implementing Lean Six Sigma into practice. Although they had a complete understanding of the company's working environment, they still faced a number of obstacles that prevented its effective implementation. Moreover, they recommended that the instrument needed to be easily usable, blend in with the SMEs environment. The preparedness self-evaluation needed to be uncomplicated, trustworthy, and resource-light, according to all of the experts.

Initial instrument creation: A preliminary list of criteria for assessing how ready a company is to apply Lean Six Sigma was created on the basis of Csfs, barriers, and existing LSS frameworks. Items were modified from previous instruments that assessed the Lean and/or Six Sigma implementation. An additional eight models (Table 4) from different sectors were also studied to extract the best practices regarding the scoring logic, design of the model, and the structure of its validation process. The draft instrument was reviewed by five industry professionals.

Table 4: Summary of the existing readiness and maturity models.

Sr.no	Authors	Focus	Inspired elements
1.	[35]	Created a self-assessment preparedness for the application of statistical procedures tailored to the food sector.	• Technique of scoring • Participation of academics and industry experts • Creation of models based on literature.
2.	[36]	Proposed a self-checking tool to assess preparedness.	• Use motivations, obstacles, and facilitators as input when creating the tool. • Completing the readiness evaluation as a team. - defining the readiness assessment's scope, such as the entire company or a particular division?
3.	[37]	Outlines a method for assessing an organization's preparedness for the fundamental ideas and components of safety change management.	• Sampling method: Developing the readiness model using Nertney's principle.
4.	[38]	Aims to examine the Swedish industry's preparedness for data analytics.	• Developing the evaluation tool using a number of preparedness capabilities. • Results of the study were used to prepare a robust model.
5.	[39]	Evaluated how well the various Industry 4.0 and smart manufacturing maturity models match the unique needs of small and medium-sized businesses.	• Rules for scoring readiness levels. - Building a solid model based on the findings of this investigation.
6.	[40]	Provide a mechanism to gauge the level of preparedness for Industry 4.0 adoption.	• Method development: Determining the level of preparedness.
7.	[34]	Proposed a maturity model for evaluating Industry 4.0 adoption	• The approach used to create the maturity model.
8.	[41]	Presented a self-assessment to help organizations measure and visualize their capabilities with respect to key elements for successful sustainability integration in the product innovation process.	• Long-term outlook on the three aspects of sustainability, with a focus on the social aspect.

Source: Authors, (2026).

Development: By updating, amending, adding, correcting, and/or eliminating some criteria and items, as necessary, this step attempted to validate the instrument. A multi-case study approach was chosen to accomplish this. This kind of approach is crucial when the subject is complicated and needs to be examined in its context [42]. Especially considering the lack of evidence on organizations' preparedness to adopt Lean Six Sigma. Employing such approach helps in the production of more trustworthy outcomes and broad conclusions. However, using this method will take a lot of time and work to complete the research.

Validation and assessment: The proposed preparedness self-evaluation model and tool were validated and evaluated throughout the fourth and final development phase. In order to reduce misunderstanding and raise the instrument's quality, queries about its functioning, item clarity, and visibility were also posed to industry specialists. An encouraging sign of the suggested tool's validity was that industrial experts said it could be used to undertake a preparedness self-evaluation for Lean Six Sigma implementation.

III.1 MATURITY LEVEL SELF-EVALUATION MODEL FOR THE IMPLEMENTATION OF LEAN SIX SIGMA

A self-evaluation maturity model has been created especially for SMEs to gauge an organization's level of readiness to apply Lean Six Sigma (LSS). The model consists of 5 components**.**

1. Commitment of top management and employees' involvement.
2. Knowledge of Tools and Techniques.
3. Infrastructure and training.
4. Motivation and Recognition
5. Social Alignment.

Each of the dimensions reflects one of the fundamental capabilities that underpin LSS implementation in SMEs, depending on which the model would be both theory-driven and practice-oriented. The following section outlines the Maturity levels, scoring mechanism, and calculation of the Lean Six Sigma Maturity Index (LSS-MI).

Table 5: Lean Six Sigma Maturity Index (LSS-MI).

Sr.no	Categories		Maturity level 1	Maturity level 2	Maturity level 3	Maturity level 4	Final Evaluation
1	**Top Management & Employees Involvement**	Does management join or attend improvement/project meetings? (e.g., project reviews, problem-solving meetings, or improvement events)?					
		Are resources (time, budget, basic tools, and key staff) allocated by top management for different projects?					
		Does top management promote a culture of learning and continuous improvement by recognizing achievements and encouraging employee contributions?					
		Are employees encouraged to share problems, ideas, or suggestions for improvement in their daily work?					
		Have employees received at least basic awareness or training on Lean, Six Sigma, or continuous improvement tools?					
2	**Infrastructure & Training**	Does the organization provide the necessary infrastructure (e.g., basic IT systems, data collection tools, process mapping tools)?					
		Are employees given awareness or basic training on Lean Six Sigma concepts and problem-solving tools?					
		Does the organization invest in training and certification for selected employees?					
		Are employees encouraged to apply learned tools (5S, DMAIC, SPC, etc.) in their daily work, with management support?					
		Does the organization have mechanisms to share knowledge from training and past projects (e.g., lessons learned, best practices)?					
3	**Tools & Techniques**	Do employees use simple quality tools (like check sheets, cause-and-effect diagrams, or basic charts) to solve everyday problems?					
		Are employees aware of Lean Six Sigma tools (e.g., Value Stream Mapping, Pareto chart, scatter plots), even if they are not using them regularly?					
		Are more structured tools and techniques (e.g., 5S, Poka Yoke, SMED, FMEA) being used in projects to improve processes?					
		Are Lean tools and some Six Sigma statistical tools (e.g., control charts, process capability, basic hypothesis					

		testing) used across most departments or teams?					
		Are advanced Lean and Six Sigma tools (both problem-solving and statistical) commonly used and integrated into daily operations?					
4	**Motivation & Recognition**	Do employees feel interested in joining improvement activities?					
		Does management say “thank you” or appreciate employees who give ideas?					
		Are small rewards (like certificates, gifts, or public praise) given for good improvement work?					
		Do employees take part in improvement work because they want to, not just for rewards?					
		Is there a regular system in place to reward and recognize employees for Lean Six Sigma efforts?					
5	**Social Performance**	Are all employees treated fairly and respectfully at work?					
		Does the company ensure that no child labor or forced labor is used?					
		Do employees have equal opportunities, without discrimination (e.g., gender, religion, caste, or background)?					
		Do employees get proper working hours, rest days, and fair treatment for overtime?					
		Does the company check workplace safety risks and take steps to prevent accidents or hazards?					

Source: Authors, (2026).

Each of the five dimensions of the maturity model is rated across four levels of maturity, reflecting the organization's capability to adopt Lean Six Sigma practices. These four levels range from minimal awareness to stable and confident deployment. The levels themselves have been adapted from classical maturity theory [43], [44] and contextualized for SMEs.

III.2 MATURITY LEVELS:

Each component is assessed across four levels of maturity, ranging from limited or uncertain adoption to full capability. The levels are described below:

Maturity Level 1: Uncertain or Limited

Organizations at this level are characterized by limited or no apparent adoption of Lean or Six Sigma. Though one or a few individuals may have former training or exposure from previous work, there is no structured implementation, no formal projects, and limited evidence of LSS-supportive behavior [45]. Resource constraints, management's lack of involvement, and low awareness among employees characteristically suppress any start [17]. This stage resembles the "Initial" level in classic maturity models in that processes are informal and at the mercy of individual effort rather than being systematically practiced [43].

Maturity Level 2: Awareness and Fair

At this stage, SMEs begin to investigate Lean and Six Sigma ideas temporarily. Basic awareness initiatives and introductory training are introduced across the departments, but systematic deployment has not been considered yet [46], [47]. In this context, organizations generally demonstrate poor standardization of processes with fragmented tools [18], [22]. This level matches the “Repeatable/Structured” phase of capability maturity, where the fundamental practices have just started to take shape but remain unstable [44].

Maturity Level 3: Understanding or Good

This level is called as ‘**Understanding or Good**’, because organizations at this level are able to comprehend and acquire new information and techniques. Compared to the previous level, this organization employs more sophisticated tools and approaches for issue resolution and has a greater understanding of LSS at the shop floor level. These organizations have moved past their own inertia in relation to LSS and have become self-learning. SMEs at this level are usually in a position to relate LSS activities to process improvements and operational performance, overcoming initial barriers including resistance from staff and absence of formal training [18],[23]. This level is an indication of increasing organizational capability and learning mindset, aligned with prevailing LSS and maturity model research.

Maturity Level 4: Capability and Certainty

This stage has been named as **'Capability and Certainty'** because organizations at this stage are better equipped to implement LSS than those at previous levels. At this stage, the organization becomes independent because the advantages are clear, and LSS begins to positively alter the organization's culture by putting in place strong infrastructure and internal training [45], [48]. Top management is involved in supporting organizational learning behaviors. SMEs in this stage generally enjoy robust leadership backing, a formal reward scheme, and wide employee engagement, allowing a self-reinforcing culture of constant improvement [17], [18]. This phase reflects the "Optimizing" stage in maturity literature, signaling institutionalized learning and strategic alignment [43].

III.3 SCORING CRITERIA:

Existing Lean Six Sigma readiness models [9],[11],[12],[23] have critical limitations; they are largely descriptive and conceptual in nature, making their practical application for SMEs limited. The present study has overcome such a limitation by adopting a scoring-based framework that converts qualitative organizational characteristics into quantifiable readiness scores. The scoring criteria are adapted from the rubric-based maturity assessment approaches commonly used in several change management and quality management models [49], [50]. Each of the 25 indicators in the proposed model is evaluated using a 0–1 scale, where 0 denotes the practice's absence, and 1 denotes its presence. This approach makes collecting evidence in SMEs easier and is consistent with the maturity evaluation's maximum score of 25. The following formula was used to determine each organization's maturity percentage:

LSS-MI= (Total Score/25) *100

The four maturity levels specified in the model were then mapped to this percentage.

Table 6: Levels of Lean Six Sigma Maturity Using Percentage Scores.

Percentage Score	Maturity Level	Interpretation
0–25%	Level 1:Uncertain or Limited	Minimal awareness; no structured LSS activity
26–50%	Level 2: Awareness and Fair	Basic awareness; inconsistent practices
51–75%	Level 3: Understanding or Good	Structured efforts; moderate tool usage
76–100%	Level 4: Capability and Certainty	Systematic, sustained LSS deployment

Source: Authors, (2026).

This structured scoring approach enhances objectivity, comparability, and practical usability for SME decision-makers.

III.4 STEPS FOR CONDUCTING SELF-ASSESSMENT.

Step 1 – Formation of the Assessment team and Model Briefing.

Performing an LSS readiness self-assessment does require the involvement of the senior management and clarity of purpose. The objective of this stage is to make the executives, managers, and owners understand why the organization is performing the assessment. A cross-functional assessment team should be assembled to perform the assessment. It should include all key operational areas represented, such as production, quality, and maintenance, among others.

Step 2: Data Collection and Preparation

Information collection by the assessment team is done based on the 25 indicators of LSS preparedness. Evidence concerning training records, tool usage, process documentation, employee involvement activities, safety practices, and leadership engagement shall be collected. All information, observations, recordings, and interviews must be impartially recorded.

Step 3: Performing Self-Evaluation for Readiness

The LSS Preparedness instrument in Table 5 is used for the self-evaluation; it has 25 indicators under five dimensions. Against each of the indicators, team members rate an entity a score of 1 (present) or 0 (absent), purely based on verifiable evidence collected in Step 3. All assessors should be asked to apply a common set of rating rules or guidelines for each item to maintain consistency and minimize subjective scoring by any assessor. Evidence may come in the form of—but is not limited to—management and employee interviews, shop-floor observations, training certificates, documented procedures, quality charts, and safety logs.

Step 4: Scoring and Determination of Maturity Level

After all the indicators have been scored, the total is summed (maximum 25). The following formula is then used to convert this score to a percentage:

LSS-MI= (Total Score)/25 *100

Based on the percentage score, each organization's level of preparedness can be determined based on the classification criteria defined in Table 6.

Step 5: Preparation of Findings and Recommendations for Action

The assessment team prepares a concise report that includes the following:

Level of maturity.
Dimension-wise performance.
Critical gaps in capacity, and
Recommended steps for addressing them.

IV. CASE STUDIES

The suggested Lean Six Sigma preparation self-evaluation technique was validated by a multiple-case study involving four SMEs in various manufacturing organizations. A LSS preparedness maturity assessment instrument was applied by each organization to determine the present maturity level through the following percentage score formula:

Percentage = (Total Score/25) × 100

A summary of these four organizations can be found in Table 7.

Table 7: Profile of Case Study Organizations.

Organization	Size	Product	Ownership	Established years
1.	SME	Manufacturer of bolts	Private	12
2.	SME	Die-manufacturing	Proprietorship	18
3	SME	Bath fitting manufacturing	Private	8
4.	SME	Automotive components	Private	19

Source: Authors, (2026).

Case 1: Bolt Manufacturing Firm

Organization 1 produces bolts for both local and export markets through semi-automated production systems. Its main products include standard fasteners, nuts, and High-Tensile Bolts. The firm scored 7/25, which is 28%, hence falling within Level 1: Uncertain or Limited. At this level, most LSS practices are absent, the infrastructure is weak, and there are only a few people with any form of LSS awareness. Top management involvement in and employee engagement with the initiative are at a minimum. Recommended improvement actions:

- Engage a consultancy firm to start basic LSS training.
- Enhance both process- and product-focused improvement activities:
- Introduce basic tools such as Scatter plots, Cause and Effect Analysis, and Pareto charts.
- Implement 5S to organize the workplace.
- Conduct periodic LSS awareness sessions for shop-floor employees.

Case 2: Die Manufacturing Firm

Organization 2 has approximately 100 employees and manufactures precision-machined parts through pressure and gravity die casting. Press tools, Sheet metal dies, forging dies, casting dies, etc., are some of the products of the company. This organization received a score of 11/25, which corresponds to 44% and positions this organization at Level 2: Awareness and Fair. The company shows initial awareness of LSS practices; however, the deployment is inconsistent and lacks a systematic structure.

Case 3: Bath Fittings Manufacturer

Organization 3 is engaged in the manufacturing of Chrome-plated bath fittings. Its products include bathroom accessories, taps, showers, etc. The company has 50 workers. The company scored 12/25, which is equivalent to 48%, and falls into Level 2: Awareness and Fair. The organization demonstrates an elementary knowledge of Six Sigma and Lean, and the application level is limited with inadequate standardization and a lack of management involvement. Improvement actions suggested for Organization 2 and Organization 3:

- Enhance the leadership and participation of the top management in the process of improvement.
- Introduce a wider set of Lean tools and basic statistical tools of Six Sigma.
- Create an official system of rewards and recognition.
- Remove organizational barriers through stronger leadership support.
- Provide external expertise to guide the LSS deployment.
- Improvement initiatives in process stability and product quality.
- Train employees in basic tools.

Case 4: Automotive Components Manufacturer

Organization 4 manufactures commercial and passenger vehicle parts. it is involved in the manufacturing of engine components, transmission parts, suspension parts, brake components, etc. It achieved the highest maturity score from the four cases, reaching 15/25 or 60%, hence placing it at Level 3: Understanding or Good. The enterprise shows stronger infrastructure, better training systems, and more concrete leadership commitment. Improvement activities are more organized, while employees show better tool awareness. Improvement actions recommended:

- Develop and document a formal project selection methodology.
- Expand the application of Six Sigma statistical methods and Lean tools.
- Institutionalize a well-structured reward and recognition system.

Cross-Case Comparison and Insights

These four cases have demonstrated that the proposed model effectively differentiates preparedness levels in SMEs for LSS: Level 1 organizations need basic awareness and infrastructure development. The organizations at Level 2 should institutionalize practices, spread training, and increase engagement at the leadership level. At Level 3, the organizations need to focus on the institutionalization of recognition systems, cross-functional learning, and the use of advanced LSS tools. By applying this maturity model, the SMEs are able to systematically evaluate their readiness, recognize capability gaps, and track their evolution into self-driven Lean Six Sigma organizations. None of the SMEs were able to reach at level 4, as this stage requires a mature organization structure, training systems, strong leadership, and frequent use of advanced tools. Typically, reaching at level 4 requires long-term investments and a sustained improvement culture, a condition that is still challenging for SMEs.

V. CONCLUSION

This research study enhances the LSS discipline by developing an integrated maturity model specifically for SMEs. The model was developed in an attempt to close the research gap, i.e., A lack of a maturity model for Lean Six Sigma, without which organizations implementing LSS cannot measure their existing level of maturity. In order to accomplish this goal, an SLR was done to explore pertinent studies on the application of Lean Six Sigma, its crucial success elements, and its hindrances. Building on these insights, the Lean Six Sigma Maturity Model was created, comprising of several maturity stages, their traits, and scores. The model was later tested and validated by interviews with subject-matter experts and subsequently applied to evaluate the current level of LSS maturity within four organizations. This model can potentially have a substantial impact on practice and knowledge. Nonetheless, it also has certain drawbacks, which are detailed in the following section, along with suggestions for additional improvement in the future.

VI. LIMITATIONS AND FUTURE DIRECTIONS FOR RESEARCH

The research done to develop the self-evaluation model/instrument has some major limitations. Comparatively few semi-structured interviews and workshops were performed. Second, in regard to the model's or instrument's use in only four organizations. Thirdly, the case studies' evaluations of social performance are restricted to the employees. It's possible that the businesses won't disclose their actual social performance regarding issues like employee maltreatment or forced labour, etc. As a result, the social performance outcomes should be interpreted with care. In order to assess whether organizations are prepared to successfully apply Lean Six Sigma, this study suggests a fresh self-evaluation model or tool. Therefore, by identifying the elements that covers a larger scope in the implementation of Lean Six Sigma, outlining problems for businesses, and suggesting the model/instrument, the study has increased knowledge in the field. It can assist practitioners in conducting a diagnostic of needs to make sure that all necessary conditions are met before beginning a Lean Six Sigma journey. It assists in the preparation phase of a Lean Six Sigma transformation, improving the likelihood that such an effort will be successful before a company invests its resources in it. This metric can be used by qualified managers to compare their organization's readiness against others because the established instrument provides a maturity score that indicates how prepared the organization is overall for launching a Lean Six Sigma initiative. The proposed methodology can also be adopted by different business activities to start their Lean Six Sigma journey, including medical, management, the hotel industry, and logistics, even though this study was done in relation to the manufacturing industry.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Sushil S Mishra and Dr. Ravi Terkar.
Methodology: Sushil S Mishra and Dr. Ravi Terkar.
Investigation: Sushil S Mishra and Dr. Ravi Terkar.
Discussion of results: Sushil S Mishra and Dr. Ravi Terkar.
Writing – Original Draft: Sushil S Mishra.
Writing – Review and Editing: Sushil S Mishra and Dr. Ravi Terkar.
Resources: Dr. Ravi Terkar.
Supervision: Dr. Ravi Terkar.
Approval of the final text: Sushil S Mishra and Dr. Ravi Terkar.

VIII. REFERENCES

[1] G. Büyüközkan, G. Kayakutlu, and I. S. Karakadılar, "Assessment of lean manufacturing effect on business performance using Bayesian belief networks," *Expert Syst. Appl.*, vol. 42, no. 19, pp. 6539–6551, 2015. doi:10.1016/j.eswa.2015.04.016.

[2] T. Kenworthy and J. Balakrishnan, "Theory usage in empirical operations management research: a review and discussion," *Manag. Decis.*, vol. 54, no. 10, pp. 2413–2432, 2016. doi:10.1108/MD-01-2016-0010.

[3] M. Godinho Filho, G. M. D. Ganga, and A. Gunasekaran, "Lean manufacturing in Brazilian small and medium enterprises," *Int. J. Prod. Res.*, vol. 54, no. 24, pp. 7523–7545, 2016. doi:10.1080/00207543.2016.1201606.

[4] S. E. Mason, C. R. Nicolay, and A. Darzi, "The use of Lean and Six Sigma methodologies in surgery," *Surgeon*, vol. 13, no. 2, pp. 91–100, Apr. 2015. doi:10.1016/j.surge.2014.08.002.

[5] J. Antony, R. Snee, and R. Hoerl, "Lean Six Sigma: yesterday, today and tomorrow," *Int. J. Qual. Rel. Manag.*, vol. 34, no. 7, pp. 1073–1093, 2017. doi:10.1108/IJQRM-03-2016-0035.

[6] J. Antony, S. Vinodh, and E. V. Gijo, *Lean Six Sigma for Small and Medium-Sized Enterprises: A Practical Guide*. Boca Raton, FL, USA: CRC Press, 2016. doi:10.1201/9781315372174.

[7] M. Dora, D. Van Goubergen, M. Kumar, A. Molnar, and X. Gellynck, "Application of lean practices in small and medium-sized food enterprises," *Br. Food J.*, vol. 116, no. 1, pp. 125–141, 2013. doi:10.1108/BFJ-05-2012-0107.

[8] S. J. Raval, R. Kant, and R. Shankar, "Revealing research trends and themes in lean six sigma," *Int. J. Lean Six Sigma*, vol. 9, no. 3, pp. 399–443, 2018.doi: 10.1108/IJLSS-03-2017-0021

[9] R. S. V. Raja and R. Raju, "A systematic literature review of Lean Six Sigma in different industries," *Int. J. Lean Six Sigma*, vol. 7, no. 4, pp. 430–466, 2016. doi:10.1108/IJLSS-12-2015-0050.

[10] G. Yadav and T. N. Desai, "Lean Six Sigma: a categorized review of the literature," *Int. J. Lean Six Sigma*, vol. 7, no. 1, pp. 2–24, 2016. doi:10.1108/IJLSS-05-2015-0015.

[11] J. Muraliraj, S. Zailani, S. Kuppusamy, and C. Santha, "Annotated methodological review of Lean Six Sigma," *Int. J. Lean Six Sigma*, vol. 9, no. 1, pp. 2–49, 2018. doi:10.1108/IJLSS-04-2017-0028.

[12] T. Ben Romdhane, A. Badreddine, and M. Sansa, "A new model to implement Six Sigma in small- and medium-sized enterprises," *International Journal of Production Research*, vol. 55, no. 15, pp. 4319–4340, 2016. doi:10.1080/00207543.2016.1249430.

[13] J. Antony, "Six Sigma vs Lean: Some perspectives," *Int. J. Prod. Perform. Manag.*, vol. 60, pp. 185–190, 2011.doi:10.1108/17410401111101494

[14] G. C. P. Condé *et al.*, "Defect reduction using DMAIC and Lean Six Sigma," *Int. J. Qual. Rel. Manag.*, ahead-of-print, 2023.

[15] G. Gokilakrishnan *et al.*, "Implementation of Lean Six Sigma techniques for tyre manufacturing in SMEs," *Int. J. Syst. Assur. Eng. Manag.*, 2023. doi:10.1007/s13198-023-01917-0.

[16] F. J. Alarcón, M. Calero, M. Á. Martín-Lara, and S. Pérez-Huertas, "Integrated Lean and Six Sigma framework," *Appl. Sci.*, vol. 14, no. 23, 2024. doi:10.3390/app142310894.

[17] W. Timans, K. Ahaus, R. van Solingen, M. Kumar, and J. Antony, "Implementation of continuous improvement based on Lean Six Sigma in SMEs," *Total Qual. Manag. Bus. Excell.*, vol. 27, no. 3–4, pp. 309–324, 2016. doi: 10.1080/14783363.2014.980140.

[18] A. Laureani and J. Antony, "Leadership and Lean Six Sigma: A systematic literature review," *Total Quality Management & Business Excellence*, vol. 30, no. 1–2, pp. 53–81, 2017.doi: 10.1080/14783363.2017.1288565.

[19] H. Araman and Y. Saleh, "A case study on implementing Lean Six Sigma," *TQM J.*, vol. 35, no. 2, pp. 337–365, 2023.

[20] A. Belhadi, F. E. Touriki, and S. El Fezazi, "Prioritizing the solutions of lean implementation in SMEs," *J. Manuf. Technol. Manag.*, vol. 28, no. 8, pp. 1115–1139, 2017.doi:10.1108/JMTM-04-2017-0066

[21] M. Kumar, K. K. Khurshid, and D. Waddell, "Status of quality management practices in manufacturing SMEs," *Int. J. Prod. Res.*, vol. 52, no. 21, pp. 6482–6495, 2014. doi:10.1080/00207543.2014.948574.

[22] A. Shokri, J. Antony, J. A. Garza-Reyes, and M. Upton, "Readiness for sustainable implementation of Lean Six Sigma," *Int. J. Qual. Rel. Manag.*, vol. 38, no. 8, pp. 1747–1770, 2021. doi:10.1108/IJQRM-08-2020-0261.

[23] A. B. Pampanelli, P. Found, and A. M. Bernardes, "A Lean & Green model for a production cell," *J. Clean. Prod.*, vol. 85, pp. 19–30, 2014. doi: 10.1016/j.jclepro.2013.06.014.

[24] M. Lande, R. L. Shrivastava, and D. Seth, "Critical success factors for lean six sigma in SMEs," *TQM J.*, vol. 28, no. 4, pp. 613–635, 2016.doi:10.1108/TQM-12-2014-0107

[25] A. B. C. Tan, D. H. van Dun, and C. P. M. Wilderom, "Lean innovation training and transformational leadership," *Int. J. Lean Six Sigma*, vol. 15, no. 8, pp. 1–31, 2024.doi:10.1108/IJLSS-06-2022-0126

[26] M. L. Cherry, *Analysis of the effect of an ergonomic improvement program*, Ph.D. dissertation, Indiana State Univ., 2019.

[27] A. Laureani and J. Antony, "Leadership – a critical success factor for Lean Six Sigma," *Total Qual. Manag. Bus. Excell.*, vol. 28, no. 3–4, pp. 394–410, 2017.doi:10.1080/14783363.2016.1211480

[28] W. Timans, J. Antony, K. Ahaus, and R. van Solingen, "Implementation of Lean Six Sigma in SMEs," *J. Oper. Res. Soc.*, vol. 63, no. 3, pp. 339–353, 2012.doi:10.1057/jors.2011.47

[29] M. Assarlind, J. Antony, I. Gremyr, and K. Bäckman, "Multi-faceted views on a Lean Six Sigma application," *Int. J. Qual. Rel. Manag.*, vol. 29, no. 1, pp. 21–30, 2012. doi:10.1108/02656711211190855.

[30] A. Cherrafi *et al.*, "Integration of lean manufacturing, Six Sigma and sustainability," *J. Clean. Prod.*, vol. 139, pp. 828–846, 2016. doi:10.1016/j.jclepro.2016.08.101.

[31] H. T. S. Caldera, C. Desha, and L. Dawes, "Exploring the role of lean thinking," *J. Clean. Prod.*, vol. 167, pp. 1546–1565, 2017, doi: 10.1016/j.jclepro.2017.05.126.

[32] R. Siegel, J. Antony, J. A. Garza-Reyes, A. Cherrafi, and B. Lameijer, "Integrated green lean approach," *J. Clean. Prod.*, vol. 240, p. 118205, 2019. doi:10.1016/j.jclepro.2019.118205.

[33] X. Fu, M. Guo, and N. Zhanwen, "Applying the green embedded lean production model," *Environ. Dev.*, vol. 24, pp. 22–35, 2017. doi:10.1016/j.envdev.2017.02.004.

[34] A. A. Wagire, R. Joshi, A. P. S. Rathore, and R. Jain, "Development of maturity model for Industry 4.0," *Prod. Plann. Control*, 2020. doi:10.1080/09537287.2020.1784482.

[35] S. A. H. Lim and J. Antony, "Statistical process control readiness in the food industry," *Trends Food Sci. Technol.*, vol. 58, pp. 133–139, 2016.doi:10.1016/j.tifs.2016.10.025

[36] C. G. Machado *et al.*, "Industry 4.0 readiness in manufacturing companies," *Procedia CIRP*, vol. 81, pp. 1113–1118, 2019. doi:10.1016/j.procir.2019.03.262.

[37] D. Levovnik and M. Gerbec, "Operational readiness for integrated management of changes," *Saf. Sci.*, vol. 107, pp. 119–129, 2018. Available: https://doi.org/10.1016/j.ssci.2018.04.006

[38] D. Gürdür, J. El-Khoury, and M. Törngren, "Data analytics readiness assessment," *Comput. Ind.*, vol. 105, pp. 153–163, 2019.Available: https://doi.org/10.1016/j.compind.2018.12.011

[39] S. Mittal, M. A. Khan, D. Romero, and T. Wuest, "Critical review of smart manufacturing and Industry 4.0 maturity models," *J. Manuf. Syst.*, vol. 49, pp. 194–214, 2018.doi:10.1016/j.jmsy.2018.10.005

[40] A. P. T. Pacchini, W. C. Lucato, F. Facchini, and G. Mummolo, "The degree of readiness for Industry 4.0," *Comput. Ind.*, vol. 113, p. 103125, 2019.doi:10.1016/j.compind.2019.103125

[41] J. Schulte and S. I. Hallstedt, "Self-assessment method for sustainability implementation," *Sustainability*, vol. 10, p. 4336, 2018.doi:10.3390/su10124336

[42] T. Hollweck, "Robert K. Yin (2014). Case Study Research: Design and Methods," *Can. J. Program Eval.*, vol. 30, no. 1, p. 108, 2016. doi:10.3138/cjpe.30.1.108.

[43] J. Becker, R. Knackstedt, and J. Pöppelbuß, "Developing maturity models for IT management," *Bus. Inf. Syst. Eng.*, vol. 1, no. 3, pp. 213–222, 2009.doi:10.1007/s12599-009-0044-5

[44] J. Pöppelbuß and M. Roeglinger, "What makes a useful maturity model?" in *Proc. ECIS*, 2011.

[45] K. Watson-Hemphill and K. N. Bradley, "Does your deployment measure up? Presenting a maturity model for Lean Six Sigma," *The Quality Management Forum*, vol. 38, no. 3, pp. 1–8.

[46] I. Boughzala and G. J. de Vreede, "A collaboration maturity model," in *Proc. 45th HICSS*, 2012, pp. 306–315, doi: 10.1109/HICSS.2012.47.

[47] P. Cronemyr and M. Danielsson, "Process Management 1-2-3," *Total Qual. Manag. Bus. Excell.*, vol. 24, no. 7–8, pp. 933–944, 2013.doi:10.1080/14783363.2013.791114.

[48] F. Voehl *et al.*, *The Lean Six Sigma Black Belt Handbook: Tools and Methods for Process Acceleration.* Boca Raton, FL, USA: CRC Press, 2013.doi:10.1201/b15163.

[49] M. S. Andrade, "Managing change—Engaging faculty in assessment opportunities," *Innov. Higher Educ.*, vol. 36, pp. 217–233, 2011.doi:10.1007/s10755-010-9169-1.

[50] F. Wilson, "The Quality Maturity Model: your roadmap to a culture of quality," *Library Manag.*, vol. 36, no. 3, pp. 258–267, 2015.doi:10.1108/LM-09-2014-0102

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 400-407. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3021

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

OPTIMIZATION OF LASER CUTTING MODES FOR PARTS MADE FROM SHEET STEEL 12X18H10T

Dmitriy Podashev[1], Anton Kisel'[2], Ivan Nikolaev[3]

[1,2,3] Kaliningrad State Technical University – Kaliningrad, Russia.

[1]http://orcid.org/0000-0001-9112-9253 ,[2]http://orcid.org/0000-0002-8014-0550 ,[3]http://orcid.org/0009-0007-1206-3489

Email: dmitrij.podashev@klgtu.ru , anton.kisel@klgtu.ru , ivan.nikolaev@klgtu.ru

ARTICLE INFO

Article History
Received: November 28, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
laser cutting,
steel sheets,
processing modes,
surface roughness,
cut geometry,
radiation frequency,
movement speed,
mode optimization

ABSTRACT

This article focuses on the investigation and optimization of laser cutting modes for sheet materials made of corrosion-resistant austenitic steel grade 12X18H10T. The main factors affecting the quality of the cut are examined, including laser power, speed of laser head movement, radiation frequency, and focal point position. A comprehensive analysis of the impact of these factors on surface roughness and the presence of visible defects on the cut edges is provided. The practical significance of this research lies in obtaining specific recommendations for optimal processing modes corresponding to various thicknesses of steel sheets. Applying these recommendations ensures achieving the required quality of processing. Implementation of the developed technology recommendations into serial production can significantly reduce the duration of the manufacturing cycle, eliminate the need for an additional stage of mechanical finishing of parts, and increase the operational efficiency of the used equipment.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Modern machine-building production is characterized by a wide variety of material processing technologies [1-4], each of which has its own area of economic and technical feasibility. Enterprises use a range of methods for manufacturing parts with complex contours from sheet materials, including mechanical machining (milling and coordinate boring operations), stamping, waterjet cutting, and laser cutting. Each method has historically established niches of application: milling provides high precision in slots and holes, stamping is economically efficient under mass-production conditions, while waterjet cutting is indispensable for materials sensitive to thermal effects. However, within the context of medium- and small-series production tasks, especially when working with alloyed steels, laser cutting appears as the most versatile and promising method. Its advantages are comprehensive in nature and have a direct impact on both economic and operational performance indicators. Among these benefits are high productivity and speed due to non-contact interaction between the tool (laser beam) and the material, eliminating costs associated with specialized cutting tools such as mills or drills that wear out and require regular replacement. Additionally, this technology does not necessitate fabrication of complicated and expensive equipment like dies and molds, significantly reducing preparation time and expenses [5-15]. Furthermore, the ability to process large-format sheet materials without pre-cutting them into smaller blanks allows optimization of layout and minimization of waste, thus improving material utilization efficiency. In Russian Federation's rocket-space industry enterprises, corrosion-resistant steel grade 12X18H10T is widely used in critical structures operating in aggressive environments. This type belongs to the class of austenitic stainless steels. The extensive product line of components made from stainless steels produced using various techniques requires high stability and quality assurance of technological processes, particularly laser cutting [16-19]. Currently, Russian enterprises commonly employ LC MSTD FO3015 laser cutting machines, often having significant service life and experiencing natural wear of their mechanical and optical components. These factors introduce uncertainty into the process since actual beam parameters and positioning accuracy may deviate from specified values.

Moreover, accompanying technical documentation either lacks specific recommendations altogether or contains only general guidelines regarding cutting modes for different materials. Specifically, there is an absence of detailed, practice-verified settings for

12X18H10T steel at varying thicknesses. As a result, operators must empirically determine cutting parameters. Consequently, results vary: acceptable quality is achieved in some cases, whereas others exhibit substantial defects. Therefore, the lack of formalized, scientifically grounded databases for cutting regimes tailored to specific materials and machinery prevents enterprises from fully realizing the potential of laser cutting technology, ensuring high-quality products, and minimizing defective output. This underscores the urgent practical necessity and relevance of targeted research efforts in this direction. The use of laser cutting machines for sheet metal processing involves addressing several technological challenges, central among which is achieving high surface quality of cuts. Current industrial practices rely heavily on subsequent finishing operations, such as manual deburring, to eliminate visible defects and bring the part up to standard condition. However, this approach introduces negative consequences: it increases overall manufacturing lead times, raises labor intensity requiring additional human resources, leading ultimately to higher final product cost. An alternative and technologically superior path would be obtaining a finished component directly during the laser cutting stage itself, avoiding any further mechanical treatment. Achieving this goal is possible only through precise determination and strict adherence to optimal laser cutting regimes unique to each “material-thickness” combination. Thus, the objective of this study is to identify and analyze adjustable technological parameters of the laser cutting process that exert a significant influence on two key metrics: the quality of the formed cut edge (absence of burr and discoloration) and compliance with required roughness specifications according to design drawings for parts manufactured from sheet material (steel 12X18H10T).

II. MATERIALS AND METHODS

Austenitic stainless steel 12X18H10T is extensively utilized in Russian rocketry because it possesses capabilities to operate under extreme mechanical loads, thermal stresses, and aggressive environments. The chemical composition of material 12X18H10T is presented in Table 1.

Table 1: Chemical composition of steel 12X18H10T.

Element	Content, %
Carbon (C)	≤ 0.12
Silicon (Si)	≤ 0.8
Manganese (Mn)	≤ 2.0
Copper (Cu)	≤ 0.3
Nickel (Ni)	9.0–11.0
Sulfur (S)	≤ 0.02
Titanium (Ti)	0.5–0.8
Phosphorus (P)	≤ 0.035
Chromium (Cr)	17.0–19.0

Source: Authors, (2026).

Steel 12X18H10T belongs to highly carbonated (from 0.6% to 1.0%) and complex-alloyed cryogenic stainless steels. Chromium and nickel serve as primary alloying elements, preventing corrosion, providing the metal with cryogenic properties (resistance to cooling) and excellent workability. Mechanical properties of steel 12X18H10T are controlled upon release of billets and are listed in Table 2.

Table 2: Mechanical properties of steel 12X18H10T.

Parameter	Value
Ultimate tensile strength	550-650 MPa
Yield strength	225-315 MPa
Density	7.92 g/cm³
Brinell hardness	170-190 HB
Elongation	35%
Fatigue limit	530 MPa

Source: Authors, (2026).

Table 3 presents analogues of steel 12X18H10T used by industries in various countries.

Table 3: Analogues of Steel 12X18H10T.

USA	AISI 321, AISI 321H, S32100, S32109
Germany	1.4541, 1.4878, X10CrNiTi18-10, X12CrNiTi18-9, X6CrNiTi18-10
Japan	SUS321
France	Z10CNT18-10, Z10CNT18-11, Z6CNT18-10, Z6CNT18-12
United Kingdom	321S31, 321S51, 321S59, LW18, LW24, X6CrNiTi18-10
Italy	X6CrNiTi18-11, X6CrNiTi18-11KG, X6CrNiTi18-11KT
China	0Cr18Ni10Ti, 0Cr18Ni11Ti, 0Cr18Ni9Ti, 1Cr18Ni11Ti, H0Cr20Ni10Ti
Bulgaria	0Ch18N10T, Ch18N12T, Ch18N9T, X6CrNiTi18-10
Hungary	H5Ti, KO36Ti, KO37Ti, X6CrNiTi18-10
Poland	0H18N10T, 1H18N10T, 1H18N12T, 1H18N9T
Czech Republic	17246, 17247, 17248
Austria	X6CrNiTi18-10KKW, X6CrNiTi18-10S
South Korea	STS321, STS321TKA, STSF321

Source: Authors, (2025).

Figure 1 shows examples of parts produced by the Russian space-rocket industry using laser cutting.

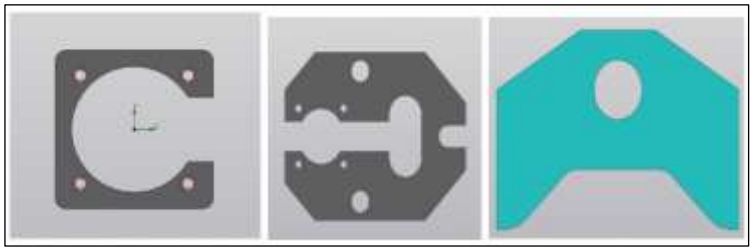


Figure 1: Parts obtained by laser cutting.
Source: Authors, (2026).

To address the task set forth in this study, we propose investigating the effect of diverse laser cutting modes and conditions on the quality characteristics of manufactured parts. The quality of the part's edge obtained via laser cutting plays a crucial role in subsequent stages of the production process [20]. For experimental studies, we employed the LC MSTD FO3015 laser cutting system (Russia), equipped with a precise coordinate system, CNC control, and an adaptive laser head featuring a sensor for monitoring clearance from the part. This setup is depicted in Figure 2.



Figure 2: Laser Cutting Machine LC MSTD FO3015.
Source: Authors, (2026).

As a result of analyzing the design documentation used at enterprises, steel sheets made from grade 12X18H10T steel were selected for experiments, which are most commonly used to manufacture parts, namely:

• Sheet 0.8 mm thick, 12X18H10T according to GOST 19903-2015/GOST 5582-75;
• Sheet 1.0 mm thick, 12X18H10T according to GOST 19903-2015/GOST 5582-75;
• Sheet 1.5 mm thick, 12X18H10T according to GOST 19903-2015/GOST 5582-75;
• Sheet 2.0 mm thick, 12X18H10T according to GOST 19903-2015/GOST 5582-75;
• Sheet 3.0 mm thick, 12X18H10T according to GOST 19903-2015/GOST 5582-75;
• Sheet 5.0 mm thick, 12X18H10T according to GOST 19903-2015/GOST 5582-75;
• Sheet 8.0 mm thick, 12X18H10T according to GOST 19903-2015/GOST 5582-75.

The test samples used were squares measuring 20×20 mm. The processing quality requirements specify surface roughness not exceeding Ra 6.3 as per the design documentation. For visual inspection and measurement of surface roughness, the following measuring instruments were employed:

1. Portable profilometer Mitutoyo Surftest SJ-410 (Figure 3).
2. Control and measurement machine KIT-MSM-32 (Figure 4).



Figure 3: Portable profilometer Mitutoyo Surftest SJ-410.
Source: Authors, (2026).



Figure 4: Control and measuring machine KIT-MSM-32.
Source: Authors, (2026).

III. RESULTS AND DISCUSSIONS

The main parameters under investigation that affect cut quality include laser power, speed of laser head movement, emitter frequency, and lens position. These parameters will be variable during the experimentation process. During the study, it was found that these parameters have an equal effect on metal thickness values of S = 0.8 mm, S = 1 mm, and S = 1.5 mm. Therefore, it was decided to combine them into one experimental group. Parameter changes occurred through the control panel of the laser machine using the Unicut software. Dependencies of surface roughness Ra on focal distance of the lens for different sheet thicknesses are shown in Figure 5.

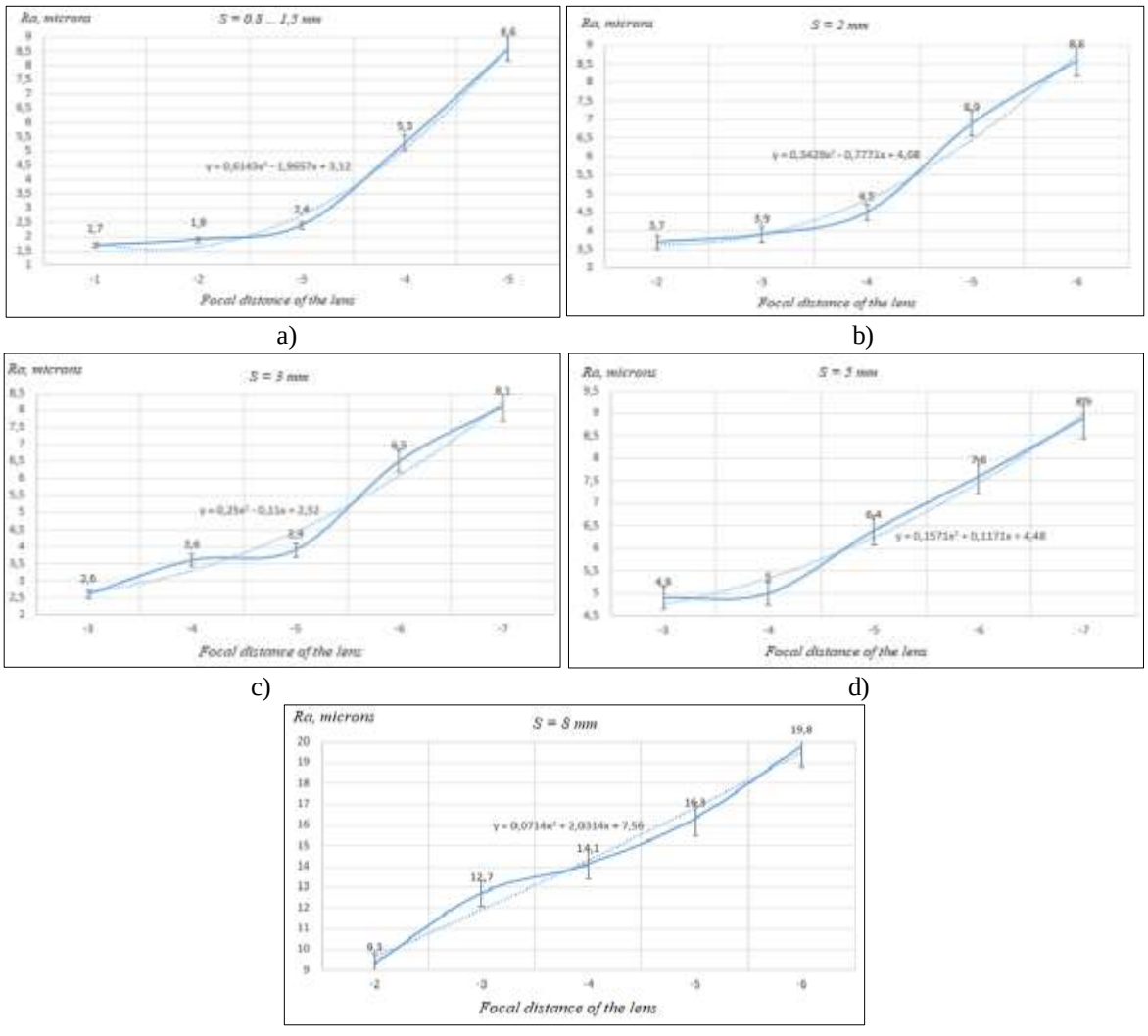


Figure 5: Dependencies of surface roughness *Ra* on lens focal distance for different sheet thicknesses:
a) for sheet thickness S = 0.8.1.5 mm; laser power P = 4000 W; laser head travel speed 7000 mm/min; frequency F = 5000 Hz;
b) for sheet thickness S = 2 mm; laser power P = 4000 W; laser head travel speed 7000 mm/min; frequency F = 5000 Hz;

c) for sheet thickness S = 3 mm; laser power P = 4000 W; laser head travel speed 7000 mm/min; frequency F = 5000 Hz;
d) For sheet thickness S = 5 mm; laser power P = 4000 W; laser head travel speed 4000 mm/min; frequency F = 1000 Hz;
e) For sheet thickness S = 8 mm; laser power P = 4000 W; laser head travel speed 2000 mm/min; frequency F = 1000 Hz.
Source: Authors, (2026).

Figure 6 shows the dependence of surface roughness Ra on the frequency of the laser emitter.

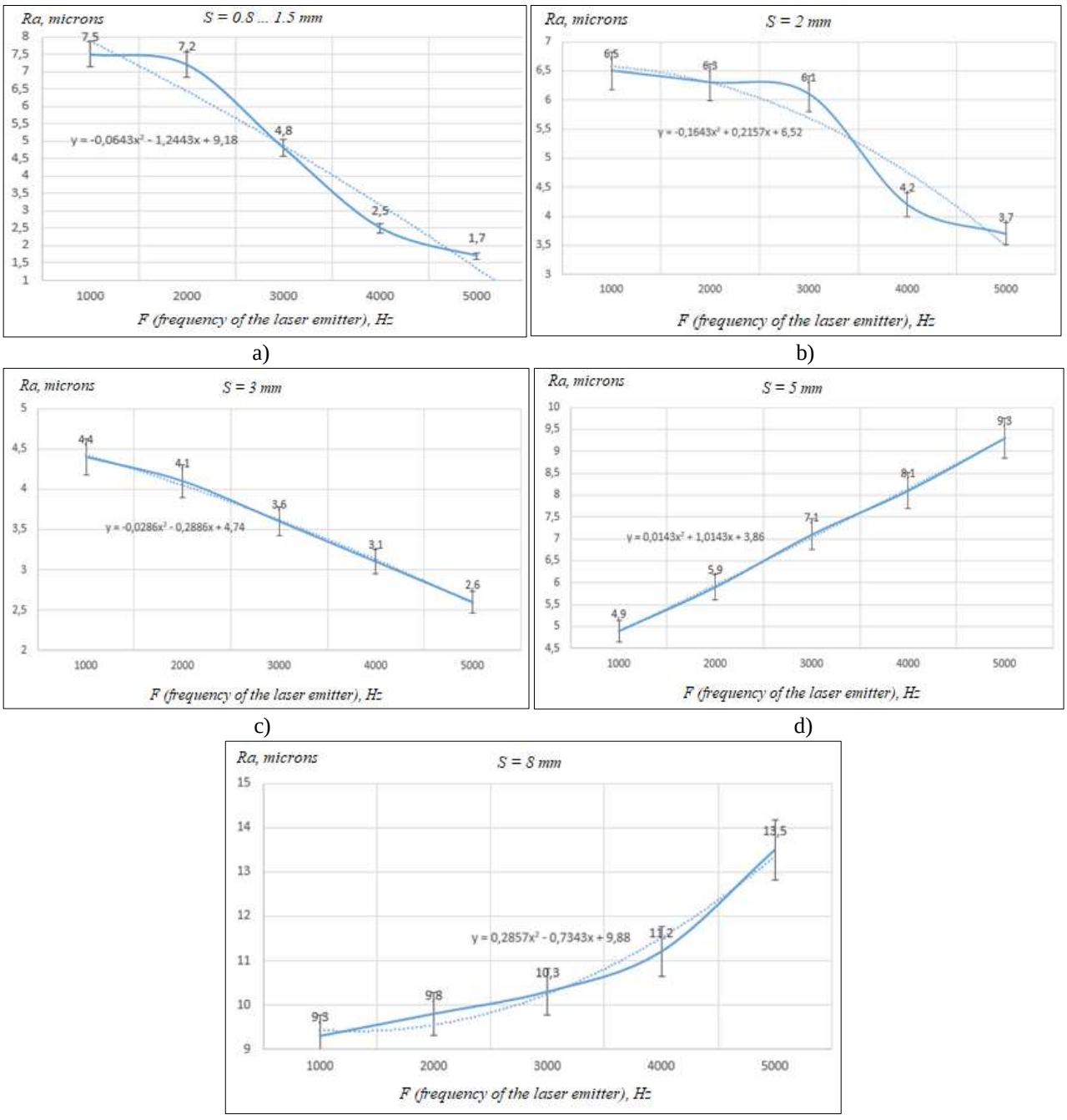


Figure 6: Dependence of surface roughness Ra on laser radiation frequency for:
a) sheet thickness S = 0.8...1.5 mm; laser power P = 4000 W; laser head speed = 7000 mm/min; lens position (focus) = -1.0;
b) sheet thickness S = 2 mm; laser power P = 4000 W; laser head speed = 7000 mm/min; lens position (focus) = -2.0;
c) sheet thickness S = 3 mm; laser power P = 4000 W; laser head speed = 7000 mm/min; lens position (focus) = -3.0;
d) sheet thickness S = 5 mm; laser power P = 4000 W; laser head speed = 4000 mm/min; lens position (focus) = -4.5;
e) sheet thickness S = 8 mm; laser power P = 4000 W; laser head speed = 2000 mm/min; lens position (focus) = -6.0
Source: Authors, (2026).

Dependences of surface roughness Ra on the speed of laser head movement for different thicknesses of cut sheets are shown in Figure 7.

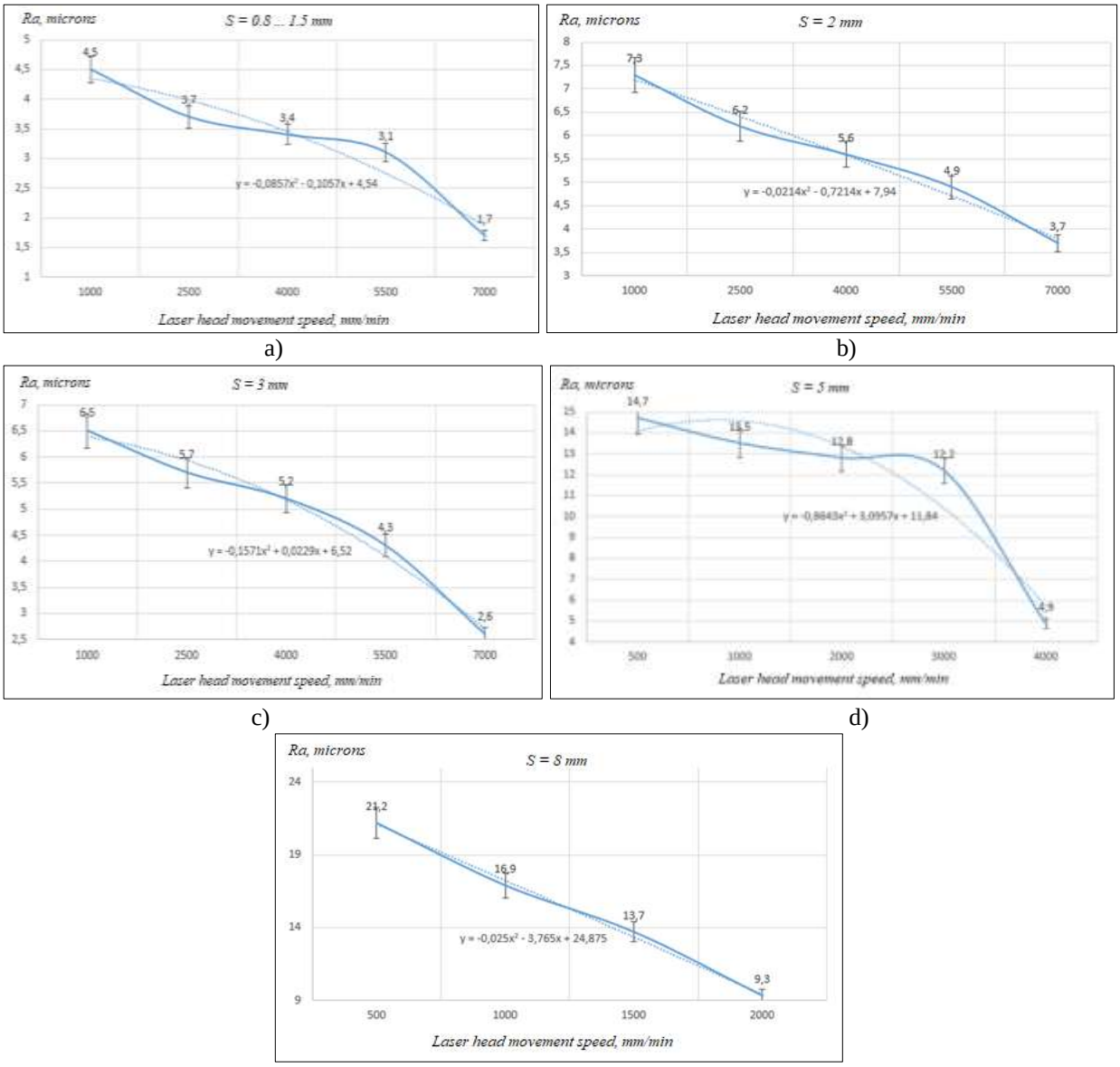


Figure 7: Dependence of surface roughness Ra on laser head movement speed at:
a) sheet thickness S = 0.8.1.5 mm; laser power P = 4000 W; laser emitter frequency = 5000 Hz; lens position (focus) = -1.0;
b) sheet thickness S = 2 mm; laser power P = 4000 W; laser emitter frequency = 5000 Hz; lens position (focus) = -2.0;
c) sheet thickness S = 3 mm; laser power P = 4000 W; laser emitter frequency = 5000 Hz; lens position (focus) = -3.0;
d) sheet thickness S = 5 mm; laser power P = 4000 W; laser emitter frequency = 1000 Hz; lens position (focus) = -4.5;
e) sheet thickness S = 8 mm; laser power P = 4000 W; laser emitter frequency = 1000 Hz; lens position (focus) = -6.0
Source: Authors, (2026).

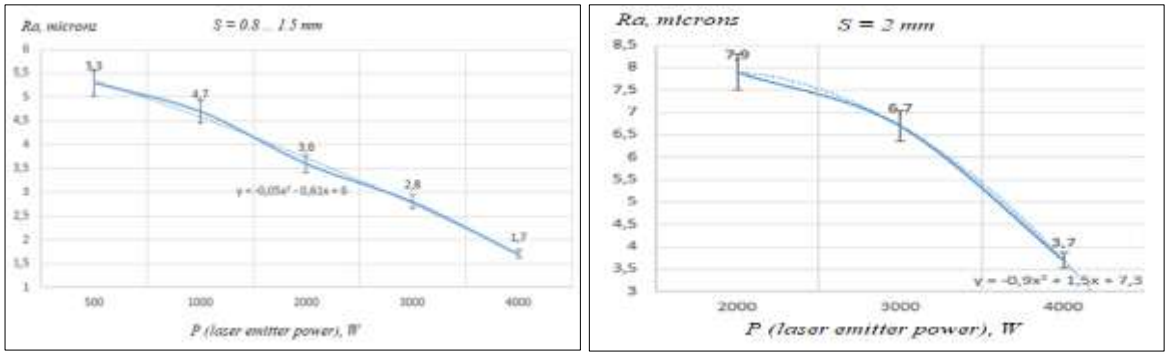


Figure 8: Dependence of surface roughness Ra on laser emitter power at:
a) sheet thickness S = 0.8.1.5 mm; laser head movement speed = 7000 mm/min; laser emitter frequency = 5000 Hz; lens position (focus) = -1.0;
b) sheet thickness S = 2 mm; laser head movement speed = 7000 mm/min; laser emitter frequency = 5000 Hz; lens position (focus) = -2.0.
Source: Authors, (2026).

In Figure 8, the dependencies of surface roughness Ra on the laser emitter power are presented. It should be noted that further experiments to reduce the laser power were not carried out when the sheet thickness was 2 mm (see Fig. 8, b), because with a further reduction in laser power, the material did not penetrate. For the same reason, no experiments were conducted to reduce the laser power for sheets thicker than 3 mm, as reducing the laser power by more than 5% resulted in insufficient penetration during one pass. As an example, Figure 9 shows images of processed surfaces after cutting sheets of 5 mm and 8 mm thickness. During visual inspection, such defects as scale formation and heat-affected zones have been identified, i.e., changes in color—from straw yellow to blue and violet along the cut line. This is particularly critical for stainless steel components since chromium can burn off in the heat-affected zones, thereby reducing corrosion resistance.

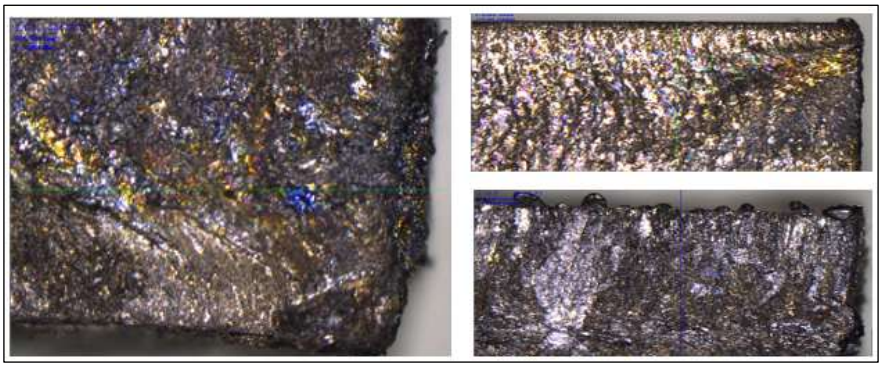


Figure 9: Images of obtained edges when cutting sheets of thickness 5 mm.
Source: Authors, (2026).

IV. CONCLUSIONS

The work performed has yielded comprehensive scientific-practical results in optimizing the modes of laser cutting of corrosion-resistant austenitic steel grade 12X18H10T using the LC MSTD FO3015 equipment. Quantitative relationships between technological parameters and quality characteristics of cuts have been established. Analysis of these relationships allows us to draw the following conclusions:

1. The optimal focus position value is found to be a critical parameter determining the geometry of the cut. For thin sheets (0.8-1.5 mm), the optimum value is -1.0 mm (surface focus), providing minimum surface roughness Ra of 1.7 microns. As the thickness increases, it becomes necessary to shift the focus deeper into the material (-4.5 mm for 5 mm, -6.0 mm for 8 mm) to ensure uniform energy input across the depth of the cut.
2. Non-linear dependencies of the optimal laser emitter power on material thickness have been revealed. For sheet thicknesses up to 2 mm, it is possible to reduce the power down to 2000 W without compromising quality, whereas for 3 mm or greater, even a slight decrease in power (to 3000 W, which represents a drop of only 25%) leads to incomplete cuts and sharp deterioration in edge quality.
3. It has been proven that increasing the speed of the laser head movement up to 7000 mm/min improves processing quality when cutting thin sheets, while for thicker ones (5-8 mm), optimal speeds range from 2000 to 4000 mm/min. Further increase in speed causes dross formation and degradation of surface quality in terms of roughness.
4. The frequency of the laser emitter exerts its most significant influence on thin sheets, where high-frequency mode (5000 Hz) ensures process stability. For thicknesses of 5-8 mm, low frequency (1000 Hz) proves optimal, promoting more effective melting and removal of material.It has been determined that a 4 kW laser emitter configured with air has the following application areas for cutting sheet materials made of 12X18H10T:

thicknesses of 0.8-2.0 mm: high-quality cutting with parameters meeting design documentation requirements;
thicknesses of 3.0-5.0 mm: limited applicability requiring subsequent mechanical treatment;
thicknesses over 5 mm: technologically unfeasible due to unsatisfactory edge quality and geometric accuracy.

For sheet thicknesses ranging from 0.8 to 2.0 mm, the recommended optimal regimes include:

laser emitter power: 4000 W;
laser head travel speed: 7000 mm/min;
frequency: 5000 Hz;
lens focal distance: -1.0...-3.0.

Under these conditions, expected surface roughness values fall within the range of Ra 1.7-3.7 microns. For thicknesses from 3.0 to 5.0 mm, recommendations include:

reducing speed to 2000-4000 mm/min;
shifting the focus to -4.5...-6.0 mm;
using pulsed mode with a frequency of 1000 Hz.

Mandatory subsequent mechanical finishing.Implementation of these recommended regimes enables elimination of manual dross removal operations, shortening of the manufacturing technology cycle, reduction of operator subjectivity, and increased utilization efficiency of the equipment.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Dmitriy Podashev, Anton Kisel, Ivan Nikolaev.
Methodology: Dmitriy Podashev, Anton Kisel, Ivan Nikolaev
Investigation: Anton Kisel.
Discussion of results: Dmitriy Podashev, Anton Kisel, Ivan Nikolaev.
Writing – Original Draft: Anton Kisel, Ivan Nikolaev.
Writing – Review and Editing: Dmitriy Podashev, Anton Kisel, Ivan Nikolaev.
Resources: Dmitriy Podashev, Anton Kisel, Ivan Nikolaev.
Supervision: Anton Kisel.
Approval of the final text: Dmitriy Podashev, Anton Kisel, Ivan Nikolaev.

VI. REFERENCES

[1] S. Tursunbaev, D. Umarova, M. Kuchkorova, A. Baydullaev, "Study of machining accuracy in ultrasonic elliptical vibration cutting of alloyed iron alloy carbon with a germanium", Journal of Physics: Conference Series, no. 1, p. 012053, 2022.

[2] J. Chen, Ya. Fan, L. Wan, W. Wu, "Surface properties of laser-woven magnesium alloys based on dot-matrix processing", Journal of Physics: Conference Series, vol. 2790, no. 1, p. 012011, 2024.

[3] C. Oliveira, R. Hamano, T. Chuvas, H. Costa, "Influence of cutting parameters on surface integrity and machining forces in top milling of metallic spraying metal coatings", ITEGAM-JETIA, no. 8(35), pp. 4-13, 2022.

[4] R. Hamano, H.R. Costa, T. Chuvas, "Evaluation of machining forces and surface integrity on AISI 304 steel top milling process under different cutting contitions", ITEGAM-JETIA, no. 5(20), pp. 166-171, 2019.

[5] Hu, Sh. "4KW Laser Tube Cutting in Metal Fabrication—TP7535M Fiber Laser Tube Cutting System", Engineering Advances, vol. 3, No. 6, pp. 505-509, 2024.

[6] V. B. Shulyatyev, "Laser-plasma cutting of aluminum-lithium alloys", AIP Conference Proceedings: ACTUAL PROBLEMS OF CONTINUUM MECHANICS: EXPERIMENT, THEORY, AND APPLICATIONS, Novosibirsk, Russia, 20–24 september 2021, vol. 2504, AIP PUBLISHING, p. 030056, 2023.

[7] P. Wen, X. Tan, J. Shan [et al.], "Numerical simulation of effect of laser power and assistant gas on laser cutting ability", Transactions of the China Welding Institution, vol. 34, no. 4, p. 57-60, 2013.

[8] H. Niino, Y. Kawaguchi, T. Sato [et al.], "Laser cutting of carbon fiber reinforced thermo-plastic (CFRTP) by IR laser irradiation", Journal of Laser Micro Nanoengineering, vol. 9, no. 2, p. 180-186, 2014.

[9] D. Rajamani, M. Siva Kumar, E. Balasubramanian, A. Tamilarasan, "Nd: YAG laser cutting of Hastelloy C276: ANFIS modeling and optimization through WOA", Materials and Manufacturing Processes, vol. 36, no. 15, pp. 1746-1760, 2021.

[10] G. Nenad, J. Meško, R. Nigrovič [et al.], "Laser cutting of ductile cast iron", Materialpruefung, vol. 62, no. 8, pp. 820-826, 2020.

[11] A. Shaparev, A. Mirzayev, M. Qarayev [et al.], "The influence gas jet on the quality laser cutting metals", Reliability: Theory & Applications, vol. 20, no. S7(83), pp. 310-319, 2025.

[12] J. Pocorni, J. Powell, J. Frostevarg, A. F. H. Kaplan, "Investigation of the piercing process in laser cutting of stainless steel", Journal of Laser Applications, vol. 29, no. 2, p.022201, 2017.

[13] R. Kristianto, C.N. Rosyidi, P. W. Laksono, "The Cutting Parameter Optimization of CO2 Laser Cutting with Taguchi-Grey Relational Analysis Method", Nanotechnology Perceptions, vol. 20, no. S11, pp. 164-174, 2024.

[14] D. J. Thomas, "The effect of laser cutting parameters on the formability of complex phase steel", The International Journal of Advanced Manufacturing Technology, vol. 64, no. 9-12, p. 1297-1311, 2013.

[15] Ya. Zhang, Ch. Wang, W. Xu [et al.], "Laser Cutting of Titanium Alloy Plates: A Review of Processing, Microstructure, and Mechanical Properties", Metals, vol. 14, no. 10, p. 1152, 2024.

[16] M. T. E. Kassim, E. T. Karash, A. M. Mahmood, Ja. N. Sultan, "Mechanical Property Enhancement of Stainless Steel 12X18H10T Through nc-TiN Coating: A Simulation Study", Annales de Chimie: Science des Materiaux, vol. 48, no. 2, pp. 163-176, 2024.

[17] A. M. Molodets, A. A. Golyshev, "Electrophysical Properties of Stainless Chromium–Nickel Steel at High Pressures and Temperatures at Stepwise Shock Compression", Physics of Metals and Metallography, vol. 125, no. 6, pp. 571-578, 2024.

[18] D. S. Makashin, A. G. Kisel', "Determination of the Range of Effective Grinding Modes and Grinding Wheel Materials for Processing 95Kh18 Steel", Russian Metallurgy (Metally), vol. 2024, no. 7, pp. 1495-1500, 2024.

[19] F. Bader, L. Bathelt, E. Djakow [et al.], "An Approach for an Innovative 3D Steel Strip Straightening Machine for Curvature and Saber Compensation", Key Engineering Materials, vol. 926, pp. 2391-2396, 2022.

[20] L. Łatka, E. Jonda, T. Kiełczawa, "The influence of the laser cutting process parameters on the quality of the cut edge", Journal of Physics: Conference Series, vol. 2676, no. 1, p. 012001, 2023.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 408-415. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3027

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

OPTIMAL CONVERSION TOPOLOGIES FOR GRID-CONNECTED PV SYSTEMS: A COMPARATIVE STUDY OF SINGLE-STAGE AND TWO-STAGE CONFIGURATIONS

Abdelmoumen Ghilani[1], Amel Terki[2], Ahmed Marouane Ghodbane[3] Zakaria Alili[4] and Oussama Belaroussi[5]

[1]LMSE Laboratory of Energy Systems Modeling, Department of Electrical Engineering, University of Biskra- Biskra 07000, Algeria
[2, 3, 5]LGEB Laboratory, Department of Electrical Engineering, University of Biskra- Biskra 07000, Algeria
[4]LEE Laboratory, University of Msila, Department of Electrical Engineering- Msila 28000, Algeria

[1]https://orcid.org/0009-0008-3522-9696, [2]https://orcid.org/0009-0001-6633-0761, [3]https://orcid.org/0009-0004-5028-116X
[4]https://orcid.org/0009-0008-3704-0897, [5]https://orcid.org/0000-0003-1777-4187

Email: abdelmoumen.ghilani@univ-biskra.dz , a.terki@univ-biskra.dz, marouane.ghodbane@univ-biskra.dz , Zakaria.alili@univ-msila.dz , oussama.belaroussi.47@gmail.com

ARTICLE INFO

Article History
Received: November 30, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Single stage,
Two stage,
PV Grid connected,
MPPT,
Incremental conductance,
Renewable Energy.

ABSTRACT

With the ongoing rise in global energy demand, grid-connected photovoltaic (PV) systems are increasingly recognised as a viable and sustainable energy alternative. These systems primarily operate using one of two distinct topologies: the two-stage (DC-DC-AC) configuration, in which the voltage from solar panels is initially boosted by a DC-DC converter before being converted to AC through an inverter. While functional, this design presents inherent drawbacks, including increased system size, higher costs, and reduced efficiency due to multiple conversion stages. To address these challenges, the single-stage (DC-AC) topology has been developed, directly linking solar panels to the inverter and eliminating the need for a DC-DC converter to enhance overall system efficiency. This study aims to identify the optimal topology for grid-connected PV systems by offering a comprehensive comparative analysis of power loss, system efficiency, and MPPT algorithm performance across these two configurations in a 150 kW grid-connected PV system under varying irradiance conditions. The simulation results underscore the superiority of the single-stage topology, demonstrating its streamlined design, cost-efficiency, and improved energy performance. Consequently, this analysis provides a valuable framework for engineers and researchers in selecting optimised configurations for grid-connected PV systems.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Due to the increasing global need for electrical energy, solar photovoltaic (PV) technology is recognised as a cost-effective solution for power generation across various sectors, including residential energy systems, small off-grid applications, and large infrastructure projects [1]. The power output of photovoltaic panels must be controlled, as it depends on external conditions such as temperature and solar irradiation. Maximum Power Point Tracking (MPPT) techniques are used to ensure that PV panels produce the maximum possible power [2]. Incremental Conductance (IncCond) and Perturb and Observe (P&O) are two of the most extensively utilised algorithms for grid-connected systems, owing to their clarity and simplicity of implementation [3]. Grid-connected PV systems use two topologies. The first is a two-stage topology, as shown in Figure 1, which uses a DC-DC converter to boost the photovoltaic voltage and track the maximum power point of the solar panels. In the second stage, a single-phase or three-phase inverter injects the electricity produced by the solar PV system into the grid and ensures its synchronization [4].

The two-stage topology has certain drawbacks, despite its benefits in controller design. Additional stages increase system complexity, which reduces overall system dependability [5]. As the number of circuit stages increases, power losses also increase, lowering the overall efficiency of energy transfer. As shown in Figure 2, a system that relies solely on the inverter is known as a single-stage topology. It is designed to eliminate the drawbacks of the two-stage topology [6].

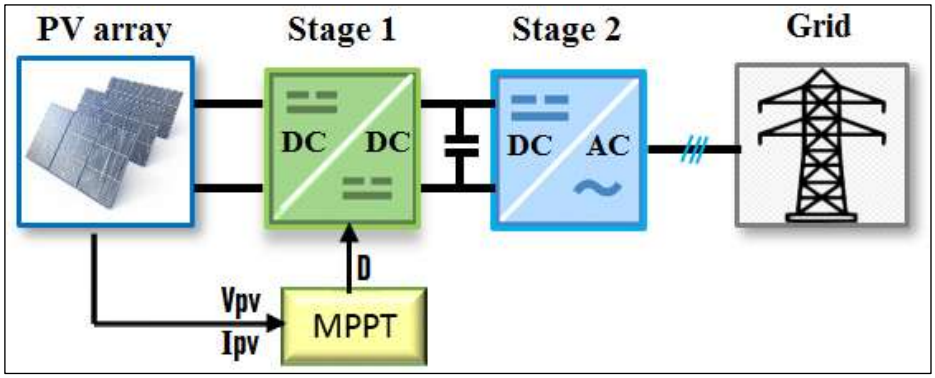


Figure 1: Two stage structure.
Source: Authors, (2026).

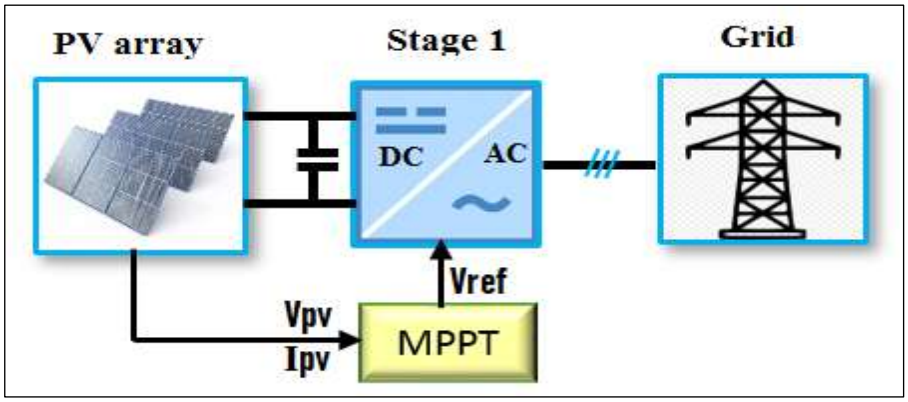


Figure 2: Single stage structure.
Source: Authors, (2026).

The primary objective of this paper is a comparative study of system efficiency, power loss, and the performance of Maximum Power Point Tracking (MPPT) methods between single-stage and two-stage topologies for a 150 kW solar photovoltaic (PV) power plant. This study provides valuable insights to support informed decision-making in the design and implementation of grid-connected PV systems. The paper is organised as follows: Section 2 details the design of the solar PV array. Section 3 discusses the modelling of the DC-DC boost converter. Section 4 describes the MPPT methods used in the analysis. In Section 5, the control technique of the grid-connected inverter is explained. Section 6 presents the simulation process and compares the results obtained. Finally, Section 7 concludes the study by summarising the key findings and their implications.

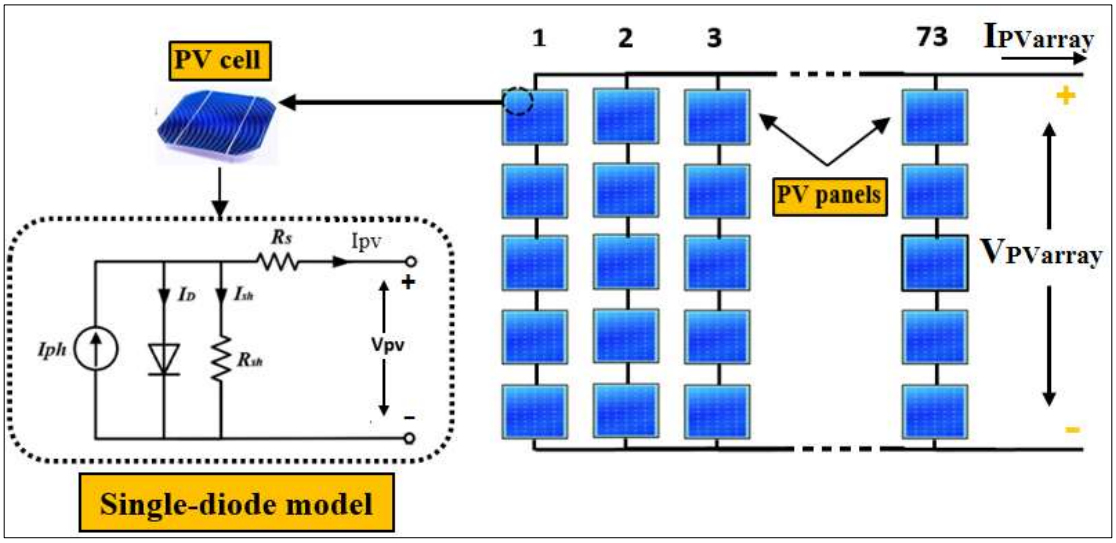


Figure 3: The PV array configuration.
Source: Authors, (2026).

II. PV POWER PLANT MODELING AND SIZING

The PV array represents the entire energy-generating system[7], consisting of multiple modules connected in series or parallel to produce high levels of current, voltage, and power. In this study, the PV system includes 365 solar panels arranged into 73 strings, with each string containing five panels. Table 1 outlines the characteristics of the PV modules under standard test conditions (STC).

Table 1: PV module characteristics.

The Parameter	The Value
Rated maximum power (P_{max})	415 W
Voltage at P_{max}	72.9 V
Current at P_{max}	5.69 A
Open-circuit PV voltage	85.3 V
Short-circuit PV current	6.09 A
The number of cells	128

Source: Authors, (2026).

Figure 3 displays the PV array design with the single-diode model, which is frequently used to study the behavior and performance of PV cells [8]. The I_{PV} can be determined by using the following equation[9]:

$$I_{PV} = I_{ph} - I_0\left[exp^{\frac{V_{pv}+I_{pv}R_s}{\alpha V_t}} - 1\right] - \frac{V_{pv} + I_{pv}R_s}{R_{sh}} \quad (1)$$

In this equation, I_{PV} represents the PV output current, and V_{PV} denotes the PV output voltage. The parameters include R_{sh} the shunt resistance, and R_s, the series resistance. I_{ph} is the current generated by solar irradiation, while V_t refers to the thermal voltage. Additionally, I_0 is the diode saturation current, and α stands for the ideality factor of the junction[3].
Figures 4 and 5 display the characteristics of the photovoltaic power plant under varying solar irradiance levels of 300, 600, 800, and 1000 W/m². These figures illustrate how irradiance affects key performance metrics, including output voltage (V_{PV}), power (P_{PV}), and current (I_{PV}).

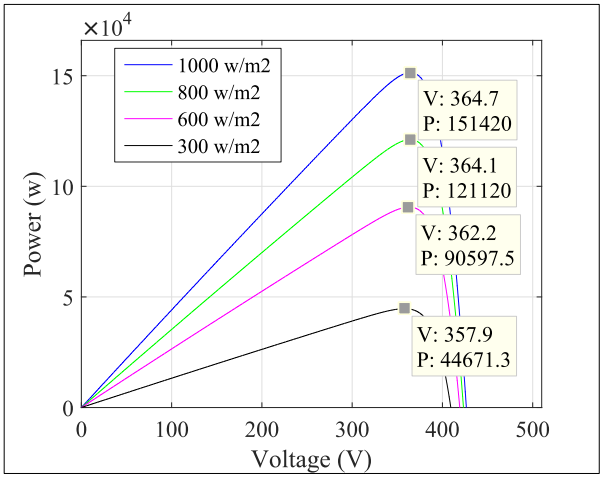


Figure 4: The power-voltage characteristic curve.
Source: Authors, (2026).

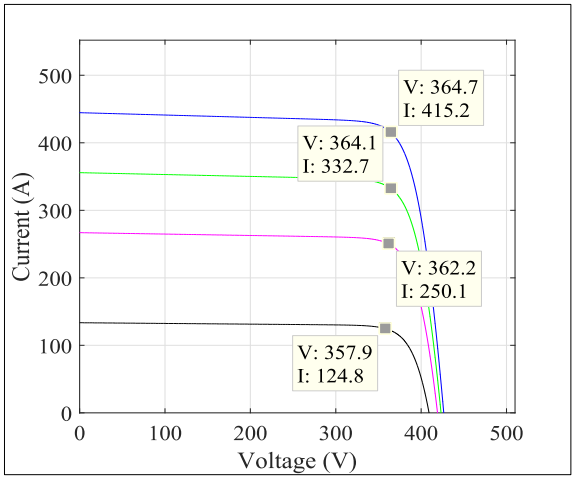


Figure 5: The current-voltage characteristic curve.
Source: Authors, (2026).

At maximum irradiance (1000 W/m²), the PV system achieves a voltage of 364.7 V with a power output of 151.42 kW, and a corresponding current of 415.2 A. As irradiance decreases to 800 W/m², the system voltage remains relatively stable at 364.1 V, but power and current reduce to 121.12 kW and 332.7 A, respectively. At 600 W/m², the voltage is 362.2 V, power output is 90.60 kW, and current is 250.1 A. Lastly, under the lowest irradiance of 300 W/m², the voltage drops slightly to 357.9 V, with power output at 44.67 kW and current at 124.8 A.

III. MODELLING OF THE DC–DC BOOST CONVERTER

The Boost Converter, as shown in Figure 6, consists of key components including an inductor, a power switch (typically a MOSFET or IGBT), a diode, and input/output capacitors for voltage filtering [10]. These components work together to boost the input voltage to a higher output level. In this system, the Maximum Power Point Tracking (MPPT) algorithm dynamically adjusts the converter's duty cycle (D) to continuously track the Maximum Power Point (MPP) of the PV array, optimizing power extraction under varying irradiance conditions [11]. The relationship between the output voltage (V_{out}) and input voltage (V_{pv}) is represented by the following equation [12]:

$$V_{out} = \frac{V_{pv}}{1 - D} \quad (2)$$

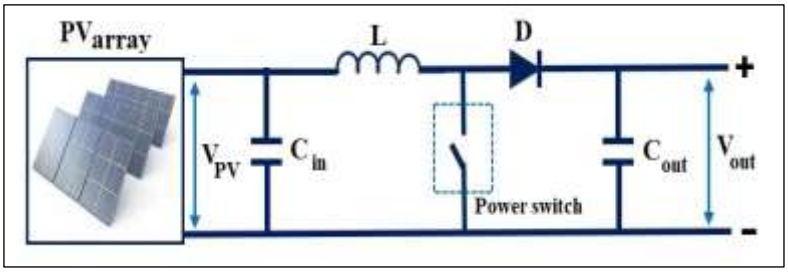


Figure 6: DC-DC boost converter.
Source: Authors, (2026).

To design the boost converter for efficient elimination of current and voltage ripples, the following equations are used [13]:

$$C_{in} = \frac{D}{8 \times f^2 \times L \times 0.1} \tag{3}$$

$$L = \frac{D(1-D)^2 \times R}{2 \times f} \tag{4}$$

$$R = \frac{P_{MPP}}{V_{grid}} \tag{5}$$

IV. INCREMENTAL CONDUCTANCE MPPT ALGORITHM

This algorithm is based on the slope of the PV system's voltage and power curvese is zero at the Maximum Power Point (MPP) [14], allowing the MPP to be tracked by comparing the incremental conductance ($\frac{\Delta I}{\Delta v}$) to the instantaneous conductance ($\frac{I}{v}$), as shown in Equation 6 [15]. This comparison indicates which side of the MPP the PV array is currently operating on[16] . Figure 7 the flowchart of the Incremental Conductance (IncCond) method.

$$\frac{dp}{dv} = \frac{dVI}{dv} = I + V\frac{\Delta I}{\Delta v} \tag{6}$$

$$\frac{\Delta I}{\Delta v} \begin{cases} = \dfrac{-I}{v} \; At\ MPP \\ > \dfrac{-I}{v} \; Left\ of\ MPP \\ < \dfrac{-I}{v} \; Right\ of\ MPP \end{cases} \tag{7}$$

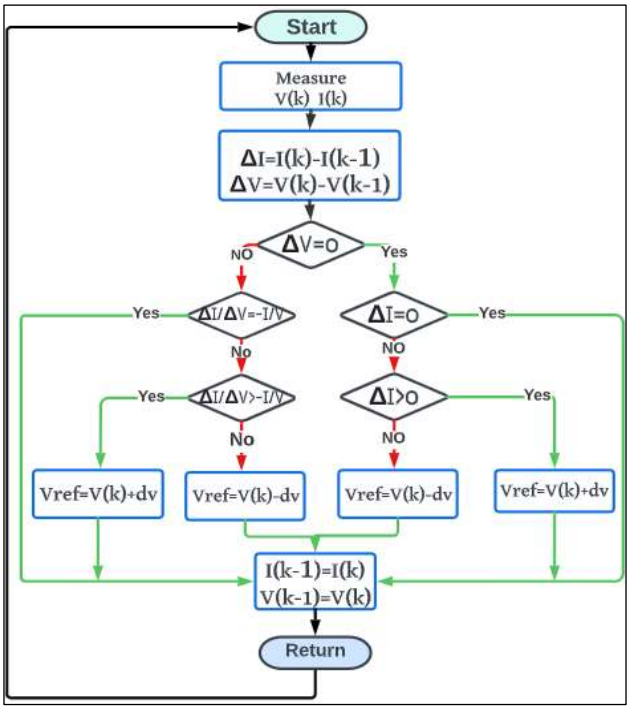


Figure 7: Flow chart of IncCond Method.
Source: Authors, (2026).

V. CONTROL STRATEGY OF PV INVERTER

The control strategy for the PV inverter consists of two main control loops. The first loop employs a proportional-integral (PI) controller to maintain the DC-link voltage at either the nominal reference voltage or the maximum power point (MPP) voltage[17], as determined by the MPPT algorithm in a single-stage configuration. The output of this loop is the reference for the active current. Meanwhile, the reactive current reference is typically set to zero to achieve a power factor close to unity[10].

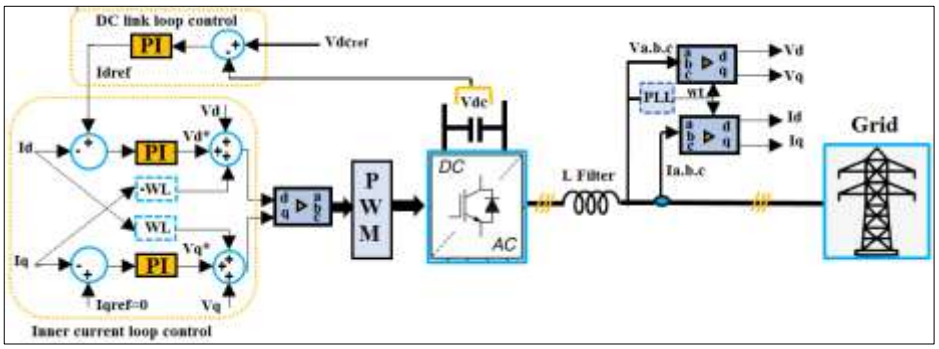


Figure 8: Control structure of the three-phase inverter.
Source: Authors, (2026).

To synchronise the inverter's output with the grid, a phase-locked loop (PLL) is used [18]. The PLL provides the phase angle required for the park transformation, which converts the grid's three-phase instantaneous voltages and currents into the dq reference frame[19]. Within this frame, inner PI controllers generate active and reactive voltage references based on the desired power output[20]. The dq-to-abc transformation then uses these active and reactive voltage references to create a reference voltage, Vabc, which is fed into the PWM generator. This generator produces the switching pulses for the inverter[4]. Figure 8 illustrates the concept of dq-control for a three-phase PV inverter.

VI. SIMULATION RESULTS AND DISCUSSION

In this section, for comparison, the performances of the two topologes. The simulations shown in Figures 9 and 10 are run, and their parameters are listed in Table 2.

Table 1: Simulation parameters.

The Parameter	The Value
U_{grid}	220 V
frequency	50 Hz
L_{Filter} , R_{Filter}	0.09 H 1.1Ω
L_{boost}	3 mH
C_{boost}	1100 µF
C_{DC}	120000 μF

Source: Authors, (2026).

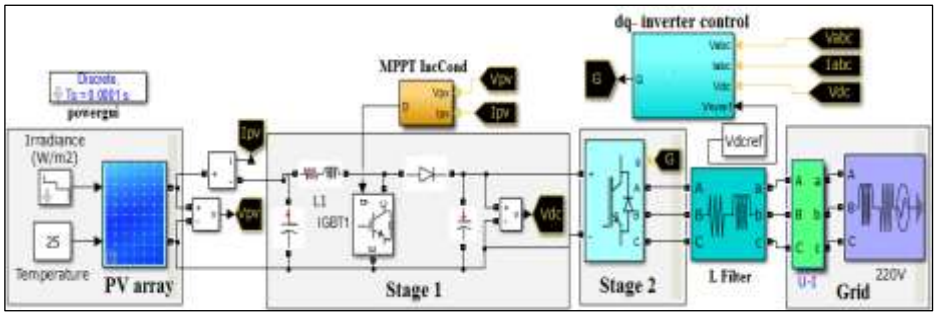


Figure 9: Simulation model of the two-stage topology.
Source: Authors, (2026).

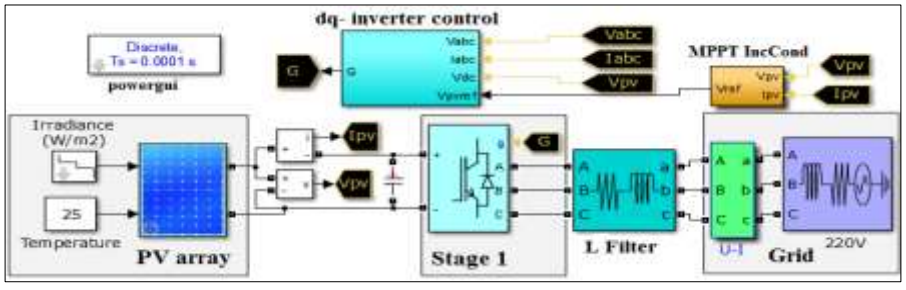


Figure 10: Simulation model of the single-stage topology.
Source: Authors, (2026).

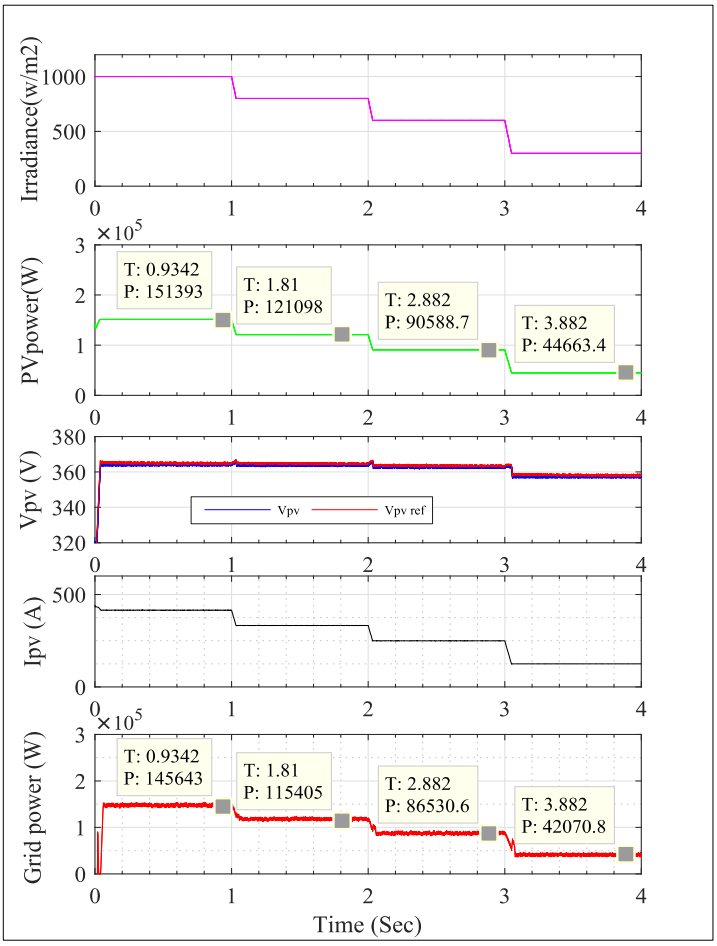


Figure 11 : Single-stage topologies Simulation result.
Source: Authors, (2026).

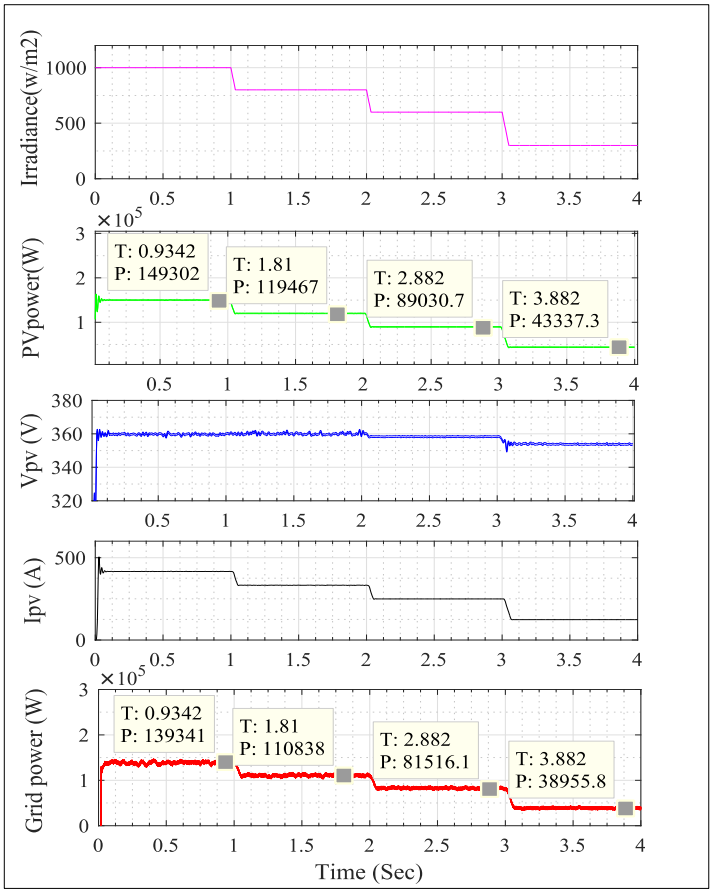


Figure 12: Two-stage topologies Simulation result.
Source: Authors, (2026).

The simulation results presented in Figures 11 and 12 illustrate key performance metrics, including voltage, current, and power output for the PV array, as well as the power injected into the grid for both single-stage and two-stage topologies. At maximum irradiance (1000 W/m²), the PV panels in the single-stage system generated an output power of 151.393 kW, with 145.643 kW successfully injected into the grid. In comparison, the two-stage system's PV panels produced a slightly lower output of 149.302 kW, of which 139.341 kW was injected into the grid due to additional conversion losses. As irradiance decreased, the single-stage topology maintained its efficiency advantage. At 800 W/m², the PV panels in the single-stage system generated 121.098 kW, with 115.405 kW reaching the grid, while the two-stage system produced 119.467 kW and injected 110.838 kW. This trend persisted at 600 W/m², with the single-stage topology producing 90.5887 kW and injecting 86.5306 kW, compared to the two-stage system's output of 89.0307 kW and grid injection of 81.5161 kW. At the lowest irradiance level (300 W/m²), the single-stage system generated 44.6634 kW, injecting 42.0708 kW into the grid, whereas the two-stage system generated 43.3373 kW with only 38.9558 kW reaching the grid.

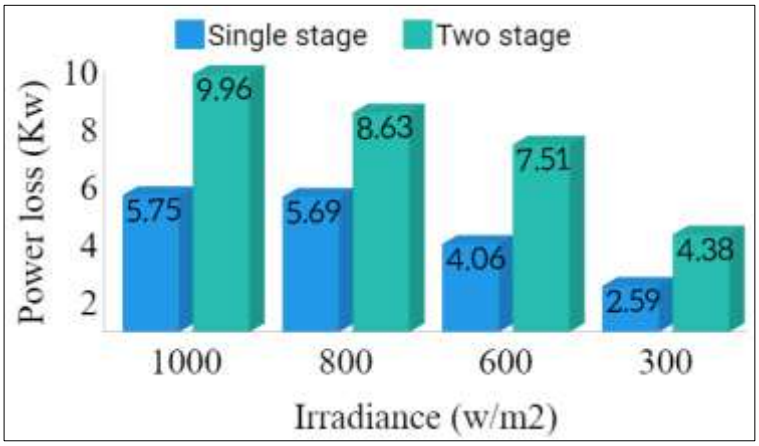


Figure 13: Quantitative Comparison of Single-Stage Two-Stage Power Losses.
Source: Authors, (2026).

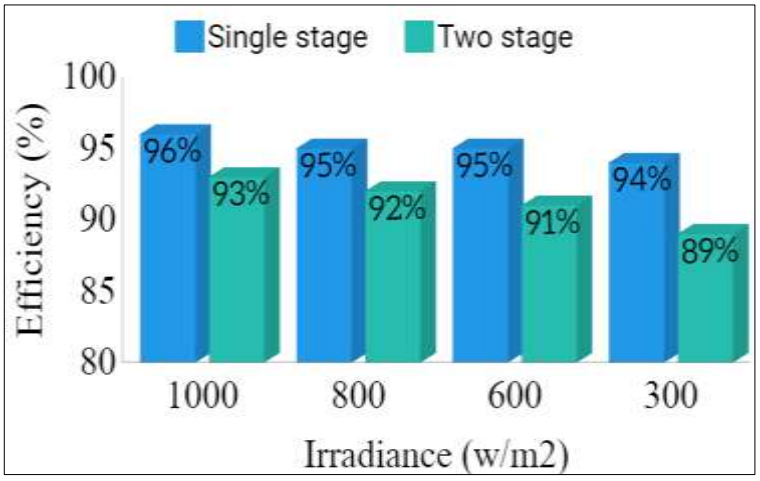


Figure 14: Quantitative Comparison of Efficiency in Single-Stage and Two-Stage Topologie.
Source: Authors, (2026).

The histogram columns in Figures 13 and 14 provide a visual, quantitative comparison between the two topologies, emphasizing the disparity in power losses at various irradiance levels. Power losses in the two-stage system reached 9.96 kW, 8.63 kW, 7.51 kW, and 4.38 kW at irradiance levels of 1000 W/m², 800 W/m², 600 W/m², and 300 W/m², respectively. In comparison, the single-stage topology showed lower losses of 5.75 kW, 5.69 kW, 4.06 kW, and 2.59 kW at the same irradiance levels. This reduction in power loss is attributed to the direct connection between the solar panels and the inverter in the single-stage design, eliminating the DC converter stage losses present in the two-stage system. The histogram effectively highlights these differences, reinforcing the efficiency advantage of the single-stage topology, which recorded efficiencies of 96%, 95%, 95%, and 94% under irradiance levels of 1000 W/m², 800 W/m², 600 W/m², and 300 W/m², respectively, as opposed to the two-stage system's efficiencies of 93%, 92%, 91%, and 89%. These results collectively support the single-stage configuration as a more efficient and economically viable choice for grid-connected PV systems, given its capacity to reduce power losses and enhance energy transfer efficiency.

VI. CONCLUSION

This study offers valuable insights into the optimisation of grid-connected photovoltaic (PV) systems through a comprehensive comparative analysis of single-stage and two-stage topologies within a 150 kW power plant framework. The findings reveal that the single-stage architecture significantly reduces power losses compared to its two-stage counterpart, resulting in enhanced overall efficiency. Furthermore, the Maximum Power Point Tracking (MPPT) algorithm exhibited superior performance in the single-stage configuration, successfully achieving a higher and more accurate maximum power point across varying solar irradiance conditions. This improvement underscores the efficacy of employing an inverter directly for maximum power tracking, as opposed to relying on a DC-

DC converter. Additionally, the single-stage topology presents multiple advantages, such as its compact size and cost-effectiveness, providing essential guidance for engineers and researchers in selecting the most optimal configurations for grid-connected PV systems.

VII. AUTHOR'S CONTRIBUTIONS

Conceptualization: Abdelmoumen Ghilani, Amel Terki, Ahmed Marouane Ghodbane, Zakaria Alili and Oussama Belaroussi.
Methodology: Abdelmoumen Ghilani, Amel Terki, Ahmed Marouane Ghodbane and Oussama Belaroussi[2]
Investigation: Abdelmoumen Ghilani, Amel Terki, Ahmed Marouane Ghodbane and Oussama Belaroussi.
Discussion of results: Abdelmoumen Ghilani, Amel Terki, Ahmed Marouane Ghodbane, Zakaria Alili and Oussama Belaroussi[2]
Writing – Original Draft: Abdelmoumen Ghilani.
Writing – Review and Editing: Abdelmoumen Ghilani and Amel Terki.
Resources: Amel Terki.
Supervision: Amel Terki, Ahmed Marouane Ghodbane, Zakaria Alili and Oussama Belaroussi.
Approval of the final text: Abdelmoumen Ghilani, Amel Terki, Ahmed Marouane Ghodbane, Zakaria Alili and Oussama Belaroussi.

VIII. REFERENCES

[1] W. V. Jahnavi and J. N. Chandra Sekhar, "A comprehensive review on application of ai algorithms for grid connected solar photovoltaic systems," Journal of Engineering and Technology for Industrial Applications, vol. 10, no. 49, pp. 87–95, Oct. 2024, doi: 10.5935/jetia.v10i49.1248.

[2] M. A. Mejias, Y. G. Landera, and L. L. Viltre, "Comparison of maximum power point tracking techniques used in photovoltaic system," Journal of Engineering and Technology for Industrial Applications, vol. 7, no. 27, pp. 4–12, Feb. 2021, doi: 10.5935/jetia.v7i27.729.

[3] Y. Houam, A. Terki, and N. Bouarroudj, "An Efficient Metaheuristic Technique to Control the Maximum Power Point of a Partially Shaded Photovoltaic System Using Crow Search Algorithm (CSA)," Journal of Electrical Engineering and Technology, vol. 16, no. 1, pp. 381–402, Jan. 2021, doi: 10.1007/s42835-020-00590-8.

[4] T. Orłowska-Kowalska, F. Blaabjerg, and J. Rodríguez, "Studies in Computational Intelligence 531 Advanced and Intelligent Control in Power Electronics and Drives." [Online]. Available: http://www.springer.com/series/7092.

[5] A. Ghilani, A. Terki, Z. Alili, A. M. Ghodbane, and O. Belaroussi, "Experimental Validation of PSO and GWO-Based MPPT for a Single-Stage Three-Phase Grid-Connected PV System Under Partial Shading," Journal Europeen des Systemes Automatises, vol. 57, no. 5, pp. 1387–1395, Oct. 2024, doi: 10.18280/jesa.570514.

[6] M. Azizi, O. Husev, and D. Vinnikov, "Single-Stage Buck–Boost Inverters: A State-of-the-Art Survey," Mar. 01, 2022, MDPI. doi: 10.3390/en15051622.

[7] A. Q. Al-Shetwi et al., "Active Power Control to Mitigate Frequency Deviations in Large-Scale Grid-Connected PV System Using Grid-Forming Single-Stage Inverters," Energies (Basel), vol. 15, no. 6, Mar. 2022, doi: 10.3390/en15062035.

[8] A. Refaat, A. E. Khalifa, M. M. Elsakka, Y. Elhenawy, A. Kalas, and M. H. Elfar, "A novel metaheuristic MPPT technique based on enhanced autonomous group Particle Swarm Optimization Algorithm to track the GMPP under partial shading conditions - Experimental validation," Energy Convers Manag, vol. 287, Jul. 2023, doi: 10.1016/j.enconman.2023.117124.

[9] M. A. Zeddini, S. Krim, and M. F. Mimouni, "Experimental validation of an advanced metaheuristic algorithm for maximum power point tracking of a shaded photovoltaic system: A comparative study between three approaches," Energy Reports, vol. 10, pp. 161–185, Nov. 2023, doi: 10.1016/j.egyr.2023.06.019.

[10] K. L. Lai, T. Thanh, and T. Tran, "Three-Phase Grid Connected Solar System With Svm Algorithm," Journal of Engineering Research and Application www.ijera.com, vol. 8, pp. 42–48, 2018, doi: 10.9790/9622-0811024248.

[11] K. Bouguerra, S. Latreche, H. Khemliche, and M. Khemliche, "Journal of Renewable Energies Comparative Study Between Synergetic Control and Sliding Mode Control using MPPT Techniques," Journal of Renewable Energies ICREPS'24 Naama, pp. 113–129, 2024, doi: 10.54966/jreen.v1i1.1281.

[12] M. H. EL-Banna, M. R. Hammad, A. I. Megahed, K. M. AboRas, A. Alkuhayli, and N. Gowtham, "On-grid optimal MPPT for fine-tuned inverter based PV system using golf optimizer considering partial shading effect," Alexandria Engineering Journal, vol. 103, pp. 180–196, Sep. 2024, doi: 10.1016/j.aej.2024.05.115.

[13] Y. Amara, R. Boukenoui, R. Bradai, and H. Salhi, "Design and Control of Two-Stage Standalone Photovoltaic Generation System," in Proceedings - International Conference on Communications and Electrical Engineering, ICCEE 2018, Institute of Electrical and Electronics Engineers Inc., Jul. 2018. doi: 10.1109/CCEE.2018.8634511.

[14] M. N. Ali, K. Mahmoud, M. Lehtonen, and M. M. F. Darwish, "An Efficient Fuzzy-Logic Based Variable-Step Incremental Conductance MPPT Method for Grid-Connected PV Systems," IEEE Access, vol. 9, pp. 26420–26430, 2021, doi: 10.1109/ACCESS.2021.3058052.

[15] V. C. Tella, B. Agili, and M. He, "Advanced MPPT Control Algorithms: A Comparative Analysis of Conventional and Intelligent Techniques with Challenges," European Journal of Electrical Engineering and Computer Science, vol. 8, no. 4, pp. 6–20, Jul. 2024, doi: 10.24018/ejece.2024.8.4.623.

[16] A. Asnil, R. Nazir, K. Krismadinata, and M. N. Sonni, "Performance Analysis of an Incremental Conductance MPPT Algorithm for Photovoltaic Systems Under Rapid Irradiance Changes," TEM Journal, vol. 13, no. 2, pp. 1087–1094, May 2024, doi: 10.18421/TEM132-23.

[17] A. Q. Al-Shetwi, M. Z. Sujod, M. A. Hannan, M. A. Abdullahd, A. S. Al-Ogaili, and K. P. Jern, "Impact of Inverter Controller-Based Grid-Connected PV System in the Power Quality," International Journal of Electrical and Electronic Engineering and Telecommunications, vol. 9, no. 6, pp. 462–469, Nov. 2020, doi: 10.18178/IJEETC.9.6.462-469.

[18] M. Mirmohammad and S. P. Azad, "Control and Stability of Grid-Forming Inverters: A Comprehensive Review," Jul. 01, 2024, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/en17133186.

[19] R. Luhtala, H. Alenius, and T. Roinila, "Practical implementation of adaptive SRF-PLL for three-phase inverters based on sensitivity function and real-time grid-impedance measurements," Energies (Basel), vol. 13, no. 5, Mar. 2020, doi: 10.3390/en13051173.

[20] I. Jamal, M. F. Elmorshedy, S. M. Dabour, E. M. Rashad, W. Xu, and D. J. Almakhles, "A Comprehensive Review of Grid-Connected PV Systems Based on Impedance Source Inverter," IEEE Access, vol. 10, pp. 89101–89123, 2022, doi: 10.1109/ACCESS.2022.3200681.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 416-427. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3028

RESEARCH ARTICLE OPEN ACCESS

EXPERIMENTAL INVESTIGATION ON MICROSTRUCTURE AND MECHANICAL PROPERTIES OF SINTERED 316 STAINLESS STEEL SPECIMEN FOR BIO IMPLANTS

Imthiyas A*[1], Bubesh Kumar D[2], Sangeetha Krishnamoorthi S[3], Vasanth kumar J[4] and Dhanush kumar[5]

[1, 2, 3, 4, 5]Department of Mechanical Engineering, Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research Foundation (DU), Tamil Nadu, India.

[1]https://orcid.org/0009-0002-1516-9426, [2]https://orcid.org/0000-0002-4308-2793, [3]https://orcid.org/0000-0003-1427-8743
[4]https://orcid.org/0009-0007-8563-6516, [5]https://orcid.org/0009-0004-8929-3379

Email: * imthiyas@avit.ac.in

ARTICLE INFO

Article History
Received: November 30, 2025
Revised: December 10, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
SS 316,
Plasma sintering,
Mechanical properties,
Biomedical,
Composites.

ABSTRACT

This project focuses on the experimental investigation of microstructural and mechanical behavior of sintered stainless steel 316 (SS316) under various processing and compositional conditions. SS316 metal powder with 99% purity and a particle size range of 30–45 microns were utilized in both advanced and conventional sintering techniques. The first phase involved Spark Plasma Sintering (SPS) at 950°C, with a pressure of 40 MPa and a dwell time of 10 minutes, producing dense and compact specimens. In the second phase, the metal powder was sintered using a conventional furnace after cold pressing at 10 kg/cm². Corrosion testing was conducted for 7 days in distilled water, seawater, and sodium hydroxide solutions to study degradation behavior. In the third phase, SS316 powder was reinforced with ceramic and non-metallic additives such as alumina, zirconia, hydroxyapatite, and silicon in different weight ratios (25–30%), followed by sintering in a muffle furnace at 900°C. This research aims to compare the microstructure, corrosion resistance, and mechanical performance of pure and reinforced SS316 under different sintering environments, offering insights into material selection for biomedical, structural, and corrosion resistant applications.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Stainless Steel 316 (SS316) is a widely used austenitic stainless steel known for its excellent corrosion resistance, mechanical strength, and biocompatibility. It is extensively applied in marine environments, medical implants, and industrial processing due to its high chromium, nickel, and molybdenum content. In powder metallurgy, sintering of metal powders enables the fabrication of complex shapes and tailored microstructures with reduced waste and energy consumption [1] [2]. Biomedical implants and medical devices have become essential components of modern health care, offering significant improvements in comfort, mobility and overall life expectancy. As medical needs continue to expand, there is a growing demand for materials that can perform safely and reliably inside the human body. The creation of such materials requires careful control over physical, chemical and mechanical characteristics, since each implant must operate under complex biological conditions [3] [4]. Advances in material science directly influence the progress of implant manufacturing, leading to the design of components that are stronger, more compatible with human tissue and capable of long-term performance. The search for improved biomaterials therefore remains a central focus in current research, particularly in the development of metals and composites that combine mechanical strength with biological acceptance [5] [6]. Spark Plasma Sintering (SPS) provides significant processing advantages such as rapid heating and sintering, high densification within a short time, and minimal grain growth, making it highly suitable for fabricating advanced metallic materials [7] [8].

SS316 is an ideal candidate for such processing due to its excellent corrosion resistance, high strength, good ductility, and inherent biocompatibility, which make it suitable for structural and biomedical applications [9] [10]. Furthermore, the incorporation of ceramic and silicon-based reinforcements enhances the performance of SS316 by improving surface hardness, increasing wear resistance, and introducing beneficial bioactivity, particularly when hydroxyapatite is used as a reinforcing phase. These combined advantages make SPS-processed and reinforced SS316 a promising material system for high-performance engineering and biomedical applications [11] [12] [13]. Powder metallurgy is a manufacturing route in which metallic powders are compacted and subsequently consolidated through controlled heating to form dense, functional components. During the sintering stage, atoms migrate across particle boundaries, promoting strong interparticle bonding, reducing voids, and improving overall mechanical integrity [14]. Owing to its capability to achieve intricate geometries and customized material characteristics, this approach has become an essential method for producing high-performance parts across various engineering sectors [15].

SS316 is valued for its excellent durability, corrosion tolerance and compatibility with biological environments. Its alloying elements, namely chromium, nickel and molybdenum, provide strong resistance to localized corrosion, particularly pitting and crevice attack in chloride rich conditions such as seawater. These characteristics make SS316 a preferred material for demanding structural and biomedical applications [16] [17]. SS316 finds widespread use in marine, chemical and biomedical environments due to its reliable performance in corrosive conditions. The material can be effectively consolidated through powder metallurgy routes, and it responds well to the addition of reinforcing phases, making it a versatile option for developing advanced composite structures [18] [19].

Spark Plasma Sintering is an advanced consolidation method in which a pulsed direct current is applied together with external pressure to compact powders quickly and efficiently. The process is known for achieving very high density within a short duration, supported by extremely fast heating rates, lower required sintering temperatures and minimal grain coarsening. These features make SPS an effective technique for producing refined and high-performance materials [20].

Previous investigations show that the use of Spark Plasma Sintering can lead to stainless steel components with finer microstructural features and enhanced mechanical properties when compared to materials produced through conventional sintering routes. This improvement is attributed to the rapid and controlled consolidation conditions provided by the SPS process [8]. Conventional sintering is carried out by heating pressed metal powders in a furnace for a prolonged duration so that atomic diffusion can occur and the particles can join together. This approach often results in enlarged grains and may not always achieve full densification, particularly when very fine powders are used. In the case of SS316, reported processing temperatures typically fall between 900°C and 1200°C, and the resulting microstructure as well as the final properties are strongly influenced by both the selected temperature and the time for which the material is held at that temperature [2]. Microstructural features in sintered materials can be examined using Scanning Electron Microscopy, which offers detailed surface imaging at very high resolution. When combined with Energy Dispersive X Ray Spectroscopy, this technique also allows the identification of the elements present in different regions of the sample.

Together, these analytical tools make it possible to observe phases, detect pores and study the distribution of constituent elements, providing valuable insight into the quality of sintering and the uniformity of composite materials [21]. Sintered SS316 exhibits mechanical characteristics such as hardness, strength and ductility, all of which depend on factors like residual porosity, grain refinement and the conditions used during consolidation. Its resistance to corrosion in chloride rich and alkaline media is one of its major advantages, although material degradation can still occur when localized attack begins or when the protective surface film loses stability. These behaviors are important for evaluating the performance of SS316 in demanding service environments [22]. Improvements in consolidation conditions along with the addition of suitable reinforcing phases can lead to noticeable gains in the mechanical and corrosion related performance of the material [23]. The addition of ceramic phases to stainless steel, including materials such as alumina, zirconia and hydroxyapatite, is known to enhance properties like hardness, resistance to wear and stability at elevated temperatures. Hydroxyapatite also provides bioactive characteristics, which makes it attractive for medical related uses.

Reports in the literature indicate that incorporating ceramic content in the range of twenty to thirty percent can improve overall performance while still retaining the corrosion resistant nature of the steel matrix [24]. Silicon can act as an additional reinforcing phase because it offers high hardness and effective heat conducting ability. Its inclusion in the metal matrix contributes to improved strength and supports better overall structural performance [25]. This study explores the sintering behavior and performance of SS316 using both advanced and conventional approaches. While traditional furnace sintering is cost-effective, SPS offers rapid densification and improved control over grain structure due to its high heating rate and applied pressure. Additionally, combining SS316 with ceramic materials such as alumina, zirconia, hydroxyapatite, and silicon may enhance its hardness, corrosion resistance, and applicability in biomedical and high-temperature environments [26]. The outcomes of this work have relevance across several industrial sectors. In the biomedical field, SS316 composites reinforced with hydroxyapatite offer potential for bone implant and orthopaedic applications due to their biocompatibility and enhanced performance.

In marine environments, the excellent corrosion resistance of SS316 makes it suitable for underwater and offshore components. Aerospace and defense applications can benefit from high-strength, wear-resistant sintered composites that contribute to lightweight structural designs. Within the energy and nuclear sectors, SS316 is already widely deployed for reactor components, and improved composite versions can further enhance resistance to stress, wear, and corrosion under demanding service conditions.

The primary objective of this study is to fabricate SS316 sintered pellets using both SPS and conventional sintering techniques and to investigate how variations in particle size and sintering parameters influence the resulting density, microstructure, and mechanical properties of SS316. The work further aims to analyze the effects of ceramic and silicon-based reinforcements on the characteristics of the sintered composites. Additionally, corrosion behaviour is evaluated by subjecting the sintered SS316 specimens to different environments, including water, seawater, and NaOH. Comprehensive microstructural characterization is carried out using Scanning Electron Microscopy (SEM) coupled with Energy Dispersive X-ray Spectroscopy (EDX), and the hardness of both pure and reinforced SS316 samples is assessed to understand their performance enhancement.

II. MATERIALS AND METHODS

The primary raw material used in this work was SS 316 powder, selected for its corrosion resistance and suitability for consolidation through powder metallurgy routes. Table 1 shows the powder specifications used in the research work. The metal powder served as the main matrix for all specimens. To prepare composite variants, several reinforcing substances were incorporated. Alumina was employed for its high hardness and chemical stability. Zirconia was selected due to its toughness and compatibility with the steel matrix. Hydroxyapatite, which contains calcium and phosphate groups, was included to introduce bioactive characteristics that are valuable for biomedical applications. Silicon powder was added as an additional strengthening phase because of its hardness and ability to assist in heat transfer during processing. Ethanol was used during the mixing stage to assist in uniform blending of the powders and to act as a temporary binder. All powders were stored in sealed containers prior to use to prevent moisture absorption and contamination.

Table 1: Powder specifications.

Property	**Value**
Purity	99%
Particle size	30-45 microns
Gross weight	270 grams
Net weight	250 grams
Density	8.0 g/cm^3
Melting point	1400-1450°C

Source: Authors (2026).

The preparation of the sintered samples followed a systematic sequence beginning with the weighing of the required quantities of SS 316 powder and the selected reinforcement materials. Table 2 represents the parameters used in the SPS study. Each powder mixture was prepared according to the intended composition, ensuring that the chosen percentages of alumina, zirconia, hydroxyapatite and silicon were accurately maintained. Ethanol was added in small amounts to aid in achieving a uniform distribution of the reinforcing particles throughout the steel matrix.

Table 2: SPS parameters used in this study.

.**Parameters**		**Value**
Sintering temperature		950°C
Applied pressure		40 MPa
Heating rate		100°C/min
Holding time		10 min

Source: Authors (2026).

The powders were blended using a ball mill operated at controlled speed to prevent excessive heat generation or unwanted structural changes in the powder particles. The fig. 1 shows the powders used in this study. The milling process continued until a visually homogeneous mixture was obtained. The blended powders were then dried at moderate temperature to remove traces of ethanol before compaction.



(a) (b) (c)

Figure 1: (a) Stainless steel 316, (b) Al_2O_3 ZrO_2 $Ca_{10}(PO_4)_6(OH)_2$, and (c) Silicon.
Source: Authors (2026).

For pellet formation, the dried powder mixtures were placed into a steel die and compacted using a hydraulic press. A constant load was applied to produce green pellets with adequate strength for handling. These compacted samples were then subjected to thermal consolidation. A portion of the specimens was processed using SPS, where current assisted heating and pressure were applied simultaneously [26]. Another portion was consolidated by the traditional furnace sintering route, where pellets were placed in a controlled atmosphere furnace and heated at the target temperature for the selected holding time. After cooling, the samples were removed, labelled and stored for testing. Figure 2 shows the flowchart of the research work.

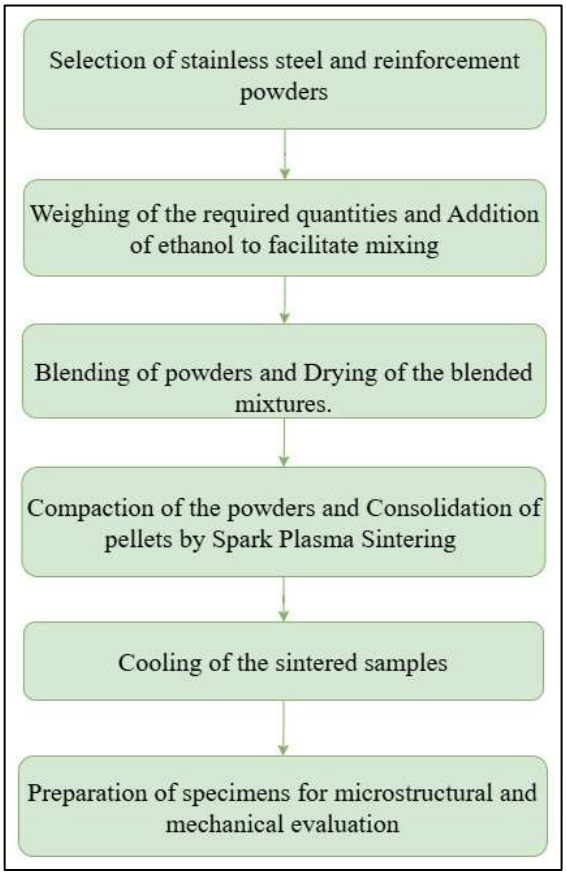


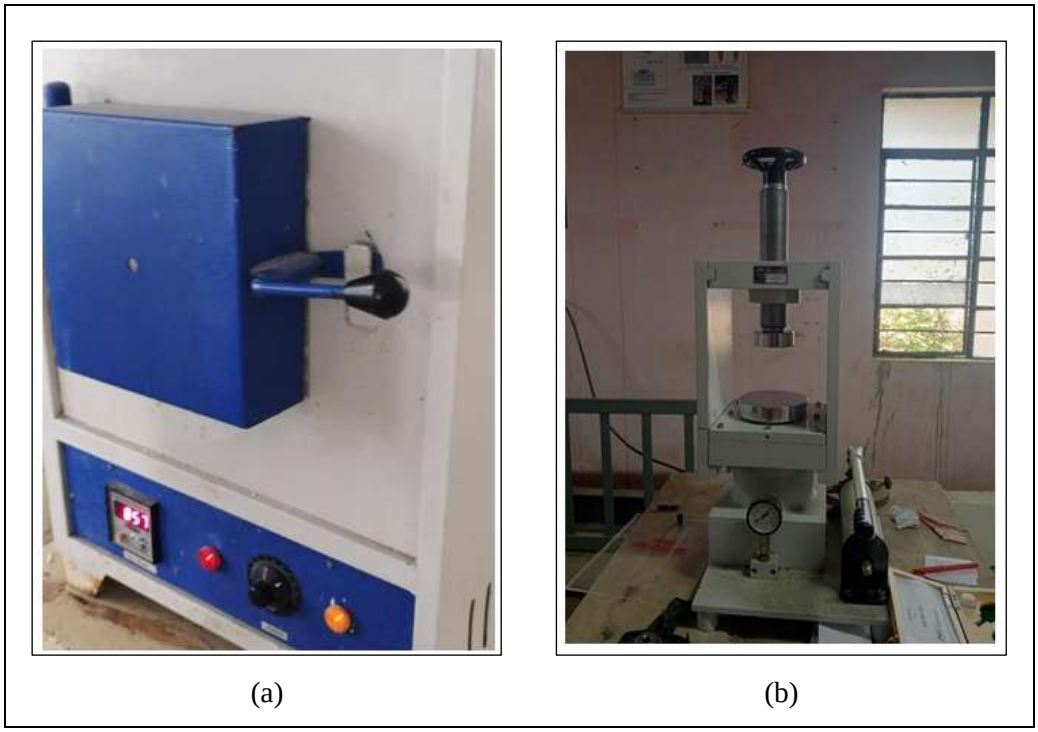
Figure 2: Flowchart of the work.
Source: Authors (2026).

A Spark Plasma Sintering system manufactured by Fuji was employed for the rapid consolidation of the prepared powders. This equipment is capable of reaching temperatures well above 1000°C and can apply pressures to 40 MPa [8]. Figure 3 represents the SPS image used for this study. The system allows the simultaneous application of electrical current and mechanical load, enabling fast heating and effective densification of the compacted samples.



Figure 3: SPS Instrument.
Source: Authors (2026).

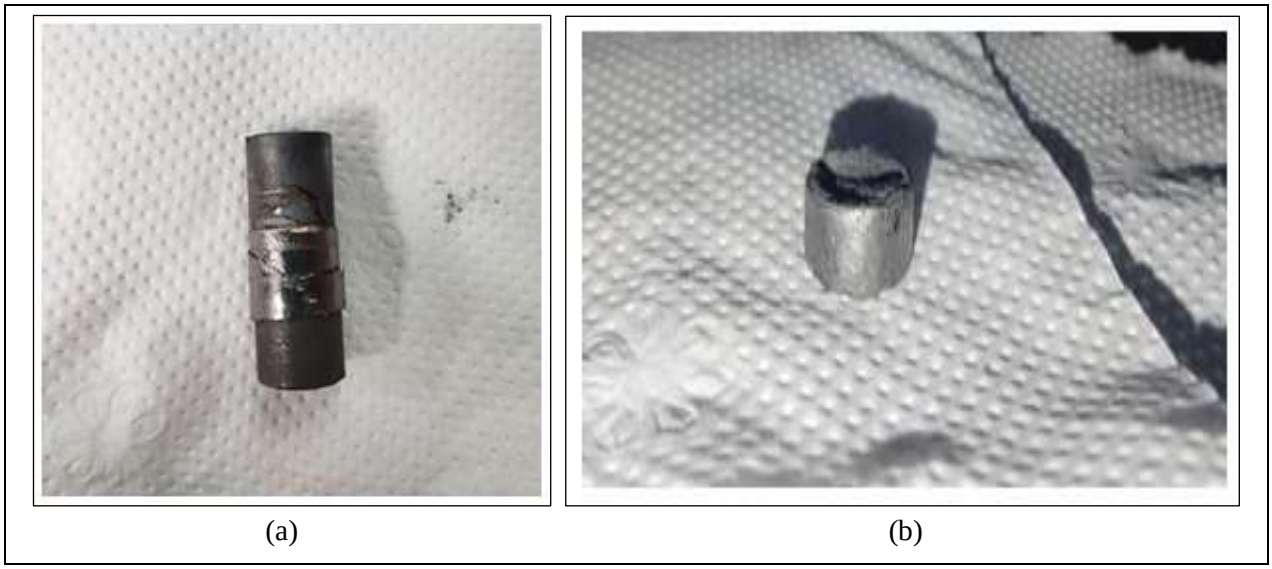
A muffle furnace was used for the conventional sintering route. Figure 4 (a), (b) represents the muffle furnace and manual pellet press used in this study. This furnace provided controlled and uniform heating for extended durations, allowing pellets to be consolidated through traditional thermal treatment. The heating chamber offered stable temperature distribution, which is essential for consistent microstructural development.



(a) (b)

Figure 4: (a) Muffle furnace and (b) Pallet press.
Source: Authors (2026).

The preparation of the pellet using SPS began with the measurement of 10g of SS 316 powder. The selected quantity was placed in a clean container and transferred into the die set that belongs to the Fuji Spark Plasma Sintering system. Figurer 5 shows the pellet made from the sintering. The sintering operation was carried out under controlled settings. These conditions were selected to promote rapid consolidation while preventing unnecessary grain growth. After the heating cycle, the pellet was allowed to cool inside the chamber. The final compact had a diameter of approximately 10 mm and a thickness of 4 mm. Once the pellet was released from the die, it was weighed, and any unutilized powder was collected for reuse.



(a) (b)

Figure 5: (a) Pallet with die and (b) Pallet.
Source: Authors (2026).

The surface of the sintered pellet was prepared to make it suitable for microstructural evaluation. A series of abrasive papers was used in sequence beginning with grade 0/3, followed by 0/2, 0/1 and finally E 120. This progression helped in removing surface irregularities and oxidation products. After the sanding stage, the specimen was transferred to a double disc polishing unit. Alumina paste was applied to the rotating surface of the polishing machine, and the pellet was gently pressed against the disc until a smooth and reflective finish was obtained. The polished specimen was rinsed thoroughly with clean water to remove any remaining abrasive particles. The resulting surface exhibited a bright mirror like appearance, which ensured clear microscopic imaging. The polished pellet produced through SPS was forwarded for microstructural and elemental examination. The SEM was used to observe pore distribution, grain boundaries and the overall surface features of the material. The same instrument was equipped with an energy dispersive X ray unit which enabled the identification of the elements present in the examined regions. This allowed verification of the chemical uniformity of the SS 316 sample. A separate set of pellets was prepared using the traditional route.

One gram of SS powder was placed inside a steel die and compressed using a manually operated press. A load of 10 g/cm2 was applied for 30 seconds to form compact samples. Three pellets were produced in this manner. Each compact had an average thickness of 3 mm and a diameter of 10 mm. The green pellets were placed inside a muffle furnace and heated to 900°C for 10 minutes. Once the heating cycle was completed, the furnace was switched off and allowed to cool naturally for 12 hours. This slow cooling reduced thermal stress in the samples. Three individual pellets were immersed in different media to evaluate their corrosion behaviour. One beaker contained distilled water, the second contained seawater, and the third contained a mixture prepared by dissolving 1.3g of sodium hydroxide in 30 mm of water. The pellets remained in these media for 8 days. Their mass and visible surface condition were recorded before and after the exposure period. Three composite formulations were prepared shown in Table 3.

Table 3: Composite sample details.

Mixture No.	Sample composition	Time Held
Mixture 1	75% SS316 + 25%	60 seconds
Mixture 2	70% SS316 + 30%	120 seconds
Mixture 3	70% SS316 + 30%	120 seconds

Source: Authors (2026).

Ethanol was added during mixing to help distribute the additives uniformly. The powders were blended manually in a grinding bowl until a consistent mixture was obtained. The prepared mixtures were pressed into pellets using the manual press. A load of 15 kg/cm2 was maintained for each pellet. The holding time was 60 seconds for the mixtures with ceramic additives and 120 seconds for the mixture that contained silicon. This ensured adequate green strength for handling [27]. All composite pellets were heated in the muffle furnace at 900°C for 15 minutes. They were then allowed to cool slowly inside the furnace for 12 hours. This natural cooling prevented cracking and maintained structural stability [20]. After processing, all composite pellets were submitted for hardness measurement. The cold compaction of powder mixtures was carried out using a manually operated pallet press. This press generated the required load to form green pellets with sufficient strength for further processing. The die set used in the press ensured that all samples were produced with identical size and shape. Surface preparation of the sintered specimens was performed using a double disc polishing machine.

This device enabled the removal of surface irregularities and produced a smooth finish suitable for microstructural examination. A series of abrasive papers was used during this process, including grades ranging from coarse to fine such as 0/3, 0/2, 0/1 and E 120. Each grade allowed gradual refinement of the sample surface before final polishing. Microstructural analysis was carried out using a scanning electron microscope equipped with an energy dispersive X ray system. The microscope provided high resolution imaging to observe features such as grain structure, pore distribution and reinforcement dispersion. The attached energy dispersive system enabled elemental analysis of selected regions, helping to confirm the presence and distribution of ceramic and silicon-based phases in the composite samples. Mechanical characterization was performed using a hardness testing machine. This instrument applied a controlled load to the polished specimen surface and measured the resulting indentation, providing an indication of the resistance of the material to localized deformation. Hardness values were used to compare pure SS pellets with those containing reinforcing phases.

III. RESULTS AND DISCUSSION

The SS 316 powder that was processed through the SPS route produced a compact pellet with clear evidence of effective consolidation. The rapid heating combined with the applied pressure resulted in strong bonding among the particles, and the final compact retained its shape without visible cracks or structural distortion. The compact had a diameter of 10 mm and a thickness of 4 mm after sintering. The recorded mass of 2.532g after the heating cycle indicated that the loss of material during the process was very small. The sintering temperature was maintained at 950°C, with an applied pressure of 40 MPa and a holding duration of 10 minutes.
These measured values point to a high degree of densification. Fig 6 represents the SEM image of the samples. The consistency in dimensions before and after processing suggests that the applied pressure and controlled temperature cycle enabled the SS 316 particles to rearrange and bond efficiently. The absence of significant mass reduction further confirms that the consolidation took place without major oxidation, evaporation or powder ejection.

Such results illustrate the capability of the SPS method to produce dense metallic pellets in a relatively short processing time when compared with conventional thermal treatments [28]. The overall outcome demonstrates that the selected sintering conditions were suitable for stainless steel grade, producing a compact with adequate density and structural stability. This provides the basis for further characterization such as microstructural analysis and hardness evaluation. The surface of the Spark Plasma Sintered SS 316 pellet was examined using SEM. The micrograph displayed a compact structure in which the particles had fused effectively, resulting in a surface with very little visible porosity. The arrangement of grains appeared consistent across the observed area, and the boundaries between adjacent grains were clearly identifiable [29]. The limited enlargement of these grains is expected, as the SPS technique employs rapid heating and a short exposure time at the peak temperature. This prevents excessive diffusion and restricts the growth of the individual grains. The surface did not show noticeable cracks, voids or defects, indicating that the consolidation process produced a stable and well bonded structure [30].

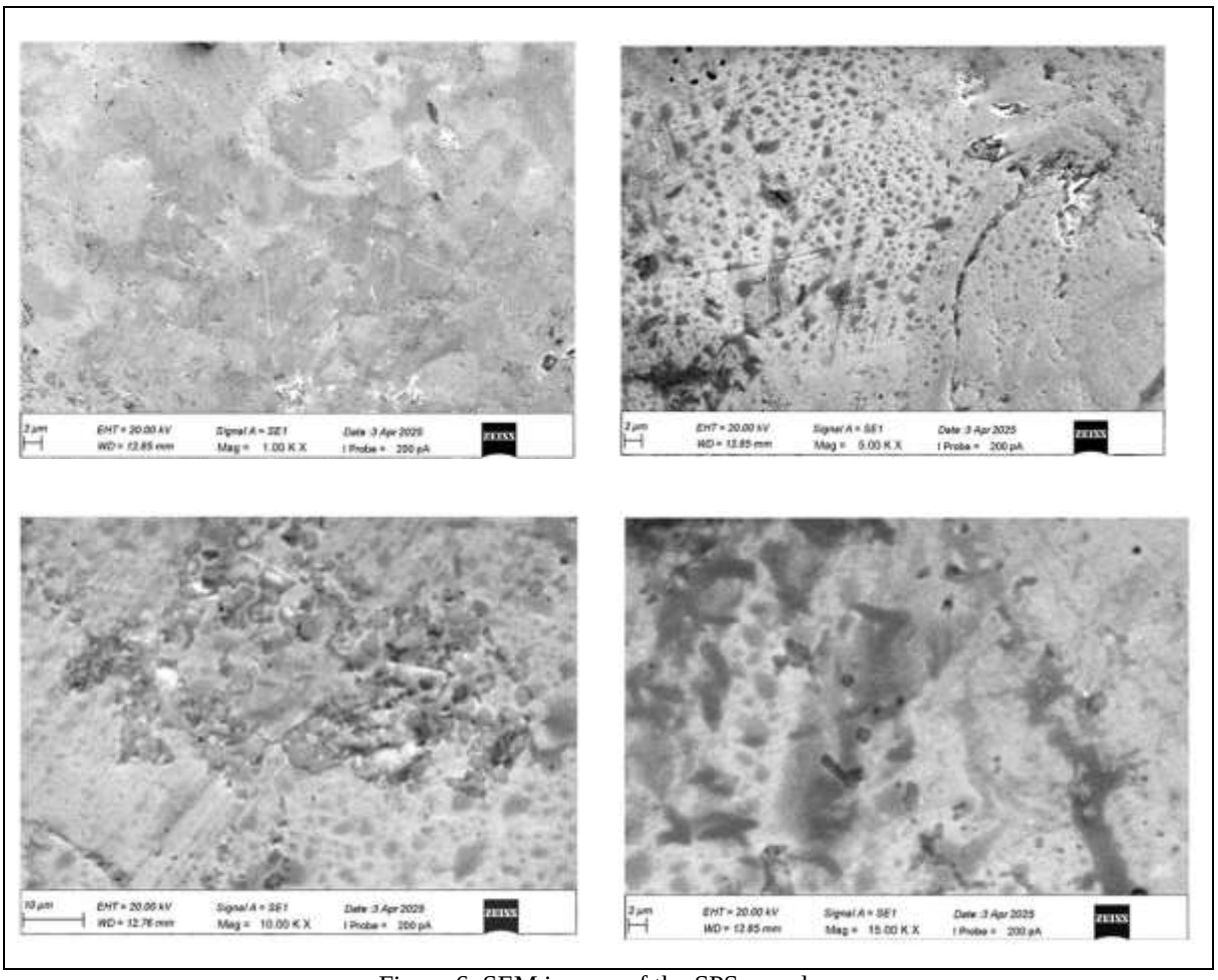
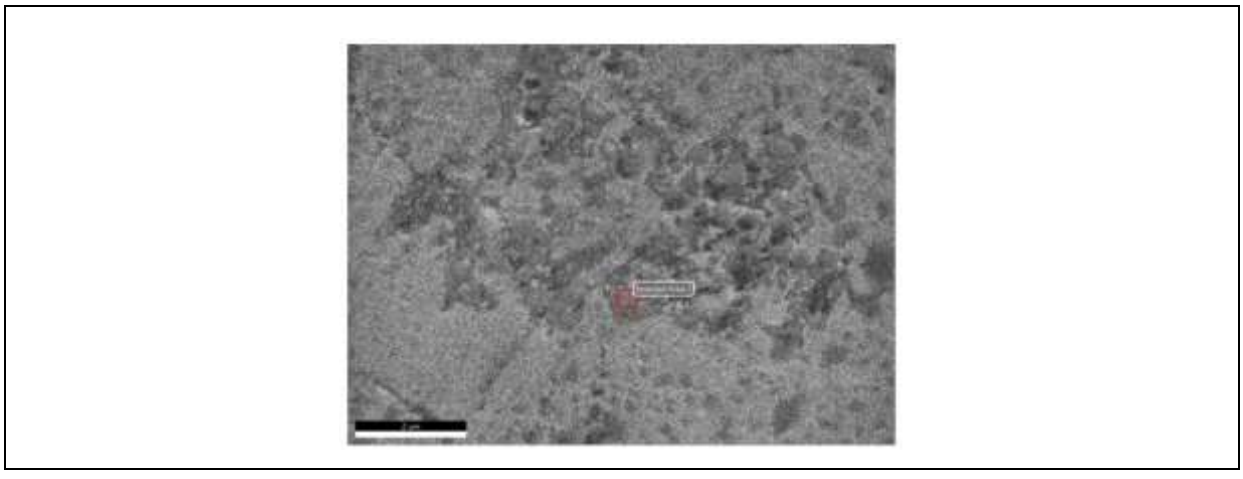


Figure 6: SEM images of the SPS samples.
Source: Authors (2026).

The microstructural features revealed through the microscopy confirm the advantages often associated with the SPS process. The presence of fine grains and the absence of large pores suggest that the material experienced efficient densification while avoiding the high degree of diffusion that typically occurs during long furnace treatments. The controlled grain size and dense structure are expected to have a positive influence on the mechanical performance of the pellet, particularly its hardness and resistance to surface deformation. Furthermore, the refined surface features and minimal oxidation indicate that the material will likely maintain favourable corrosion behaviour when exposed to aggressive environments [31]. Overall, the observed microstructure supports the conclusion that the chosen processing conditions were effective for producing a well consolidated SS sample.

Energy dispersive X ray examination was carried out at two selected regions of the Spark Plasma Sintered pellet. Figure 7 and 8 represents the EDX area 1 and 2 with the elemental mapping. The spectra from both regions presented clear and well resolved peaks corresponding to the elements that are normally present in stainless steel grade three hundred and sixteen. In the first region, elements such as carbon, oxygen, aluminium, silicon and titanium were detected in addition to the major constituents chromium, manganese, iron and nickel [1]. The presence of oxygen is attributed to slight surface oxidation during polishing, which is commonly observed in stainless steels when exposed to air. The second region displayed a composition dominated by chromium and iron, with smaller contributions from carbon and manganese.



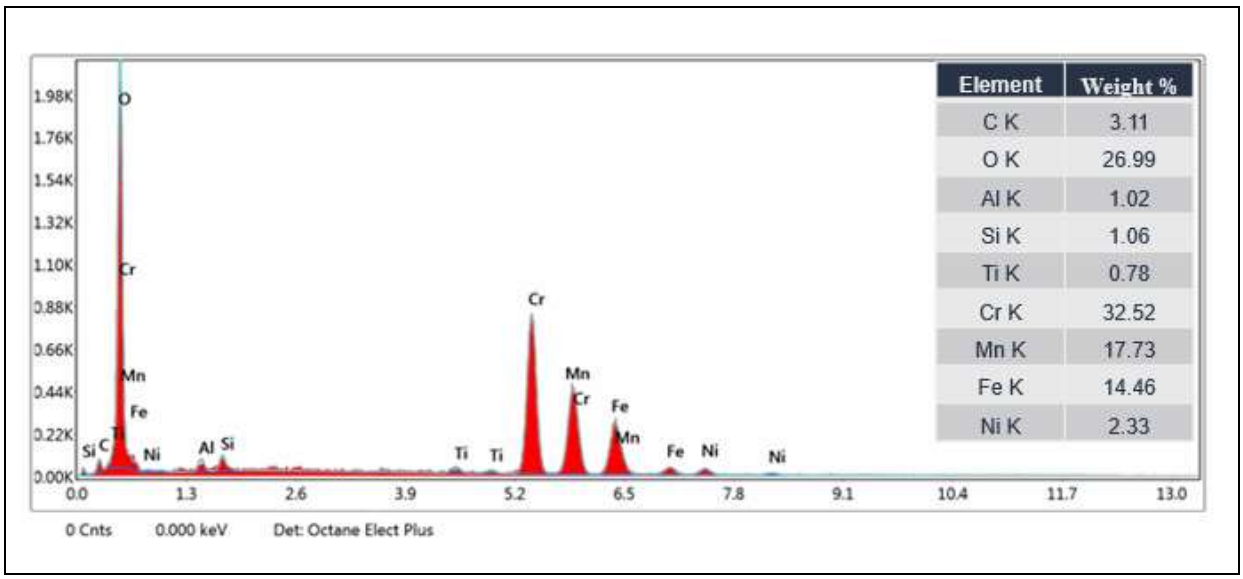


Figure 7: EDX area 1.
Source: Authors (2026).

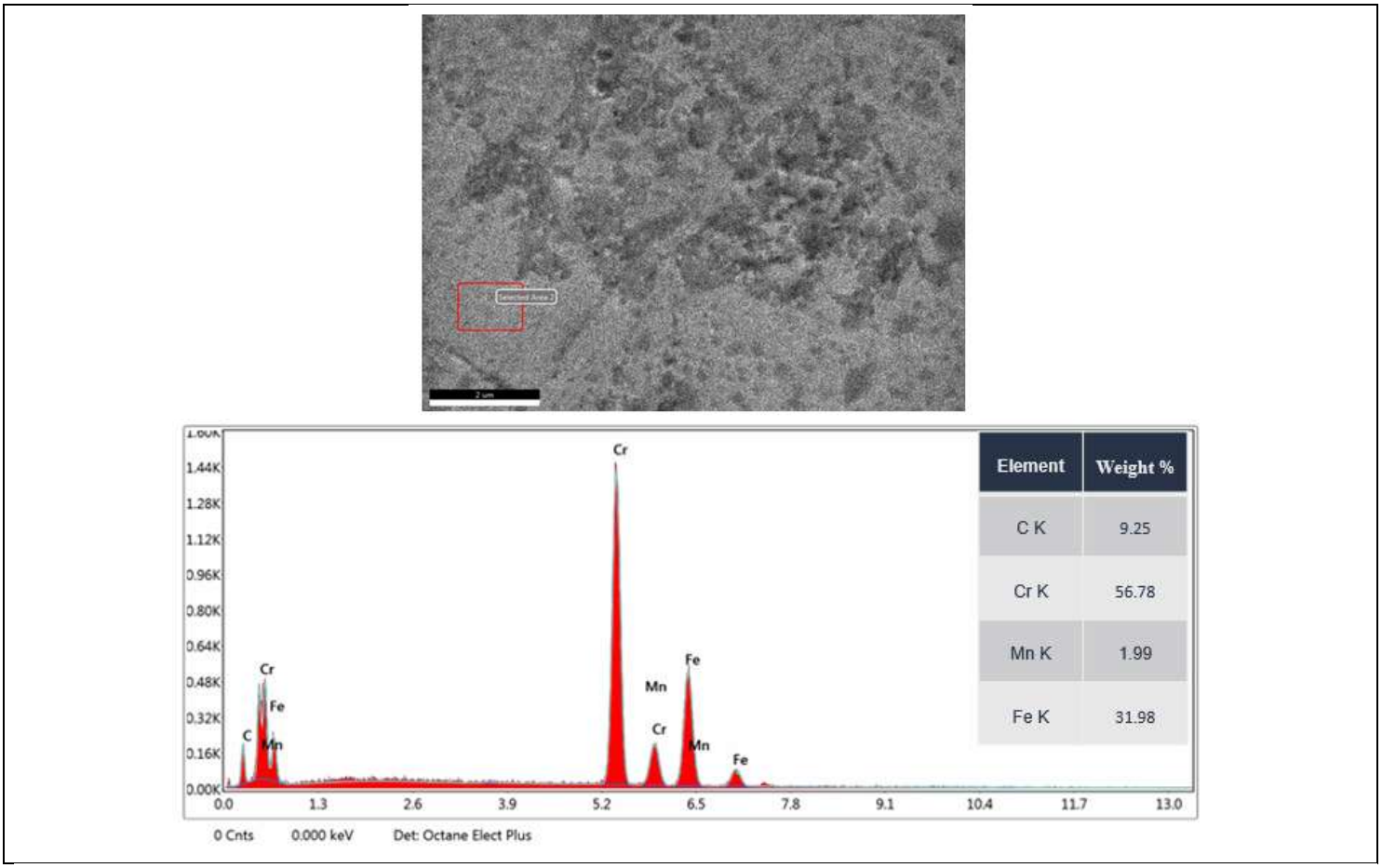


Figure 8: EDX area 2.
Source: Authors (2026).

The elemental concentrations observed in both regions fall within the expected range for stainless steel grade three hundred and sixteen. The detection of chromium, manganese, iron and nickel in appropriate proportions confirms that the alloy composition remained stable throughout the sintering process. The occasional presence of aluminum, silicon and titanium in the first region is attributed to minor traces left during surface preparation and does not indicate contamination of the bulk material [2]. The consistency in the distribution of the primary alloying elements across the examined regions demonstrates that the SPS method facilitated proper mixing and retention of the chemical composition. The process conditions prevented oxidation of the internal structure and avoided any loss of metallic constituents. The uniform chemical profile also supports the microstructural observations, showing that the rapid heating and controlled atmosphere of the sintering technique help maintain the integrity of the stainless steel powder during consolidation [7]. The combined results from both microstructural and elemental evaluation confirm that the sintered pellet preserved its intended composition.

No foreign substances or undesirable phases were detected, showing that the preparation and processing steps were performed without introducing impurities. The even distribution of chromium and iron, in particular, indicates that the corrosion resistant nature of the alloy is expected to be maintained in the final compact. These findings reinforce the reliability of the SPS route for producing dense SS 316 samples with stable chemical characteristics. Three pellets were produced through cold pressing followed by furnace sintering. All samples had a diameter of 10 mm and a thickness of 3 mm after compaction. Their appearance after the thermal cycle revealed slight variations that reflect the influence of traditional sintering conditions [23]. The first pellet maintained a stable shape and displayed a faint brown surface tint. This colour change is commonly associated with mild surface oxidation that forms during heating in a muffle furnace. The second pellet showed a very small reduction in size, suggesting a moderate level of shrinkage due to particle rearrangement and bonding during the heating stage. The third pellet exhibited a slightly uneven surface, indicating that the particle fusion was not completely uniform. Such surface roughness is typical when fine SS 316 powders are sintered without applied pressure [32].

The long cooling period of nearly 12 hours inside the furnace allowed the samples to return to room temperature gradually. This slow cooling helped reduce thermal stresses and promoted steady phase development within the material. It also prevented cracking, which is sometimes observed when metallic pellets are cooled abruptly. The visual and dimensional characteristics of the pellets confirm the expected behaviour of SS 316 during conventional furnace sintering. Compared with the Spark Plasma Sintered pellet, these samples show evidence of lower densification, as indicated by minor shrinkage and surface irregularities [33]. The appearance of oxidation on the surface also aligns with the furnace atmosphere, which does not offer the same level of control as vacuum based sintering processes. Overall, the results demonstrate that while conventional sintering can produce reasonably compact pellets, the degree of densification and surface quality is influenced strongly by furnace atmosphere, heating duration and absence of external pressure. These observations highlight the contrast between rapid current assisted consolidation and traditional thermal processing.

III.1 CORROSION BEHAVIOUR

The pellets produced through the conventional sintering route were immersed in three different media to examine their corrosion behaviour. The weights recorded before and after the 8-day exposure period is listed in Table 4. All samples showed a measurable reduction in mass, although the extent of loss differed depending on the environment.

Table 4: variation of weights.

Environment	Initial weight (g)	Final weight (g)	Change (g)	Remarks
Distilled Water	58.598	52.246	6.352	Minor expected
Sea Water	62.278	56.455	5.823	Possible pitting
NaOH Solution	65.911	61.181	4.73	Alkaline corrosion

Source: Authors (2026).

The pellet placed in distilled water lost 6.352 grams during the test. Distilled water contains no dissolved salts or aggressive ions, so the weight loss is attributed mainly to mild surface reactions and the removal of loosely bonded oxides. The pellet exposed to seawater lost 5.823 grams. Although the total mass loss is slightly lower compared with distilled water, the surface of this pellet is expected to show signs of localized attack caused by chloride ions. The sample immersed in the sodium hydroxide solution lost 4.73 grams. The alkaline nature of the medium often leads to the formation of surface films that can either protect or deteriorate depending on the material [34]. The reduction in mass in this case indicates that the film was unable to fully prevent the reaction between the pellet and the solution. The general trend observed in the weight measurements aligns with the expected behaviour of SS 316. The sample in distilled water experienced only mild degradation, confirming that SS 316 is largely stable in neutral environments. The seawater sample was subjected to chloride ions, which are known to disrupt the passive oxide layer that protects stainless steel.

This can lead to localized pits even if overall mass loss appears moderate [35]. The presence of such ions makes seawater a more aggressive medium despite the slightly lower numerical weight change [20] [36] . The specimen in the sodium hydroxide solution reacted with the alkaline medium. Although SS 316 shows some resistance in alkaline solutions, prolonged contact can still weaken the passive film and produce surface dullness or localized thinning. The mass loss recorded for this pellet indicates that the reaction rate was not negligible. These findings demonstrate that the corrosion resistance of the conventionally sintered pellets varies significantly with the nature of the environment. Neutral water causes minimal damage, seawater introduces the possibility of pitting through chloride action and the sodium hydroxide solution encourages surface layer disruption. The results also highlight the influence of sintering method on corrosion performance, since pellets produced without external pressure and rapid heating tend to contain surface irregularities that can further accelerate attack.

III.2 COMPOSITE SAMPLES OF REINFORCED STAINLESS STEEL

Three composite pellets were produced by introducing different reinforcing phases into stainless steel. The first composite, labelled C1, contained 75 % SS and 25 % of a combined mixture of alumina, zirconia and hydroxyapatite. This sample developed a uniform surface after sintering and showed no visible cracking. The second composite, labelled C2, contained 70 % SS and 30 % SS of the same ceramic mixture. This pellet also displayed a smooth surface but showed signs of increased rigidity due to the higher ceramic content. The third composite, labelled C3, was prepared with 70 % SS and 30 % silicon powder as shown in Table 5. This sample appeared dense but revealed a slightly brittle character after sintering, which is consistent with the presence of silicon rich phases that tend to limit ductility [27] [37] [38].

Table 5: Composite weight percentage with sintered temperature.

Sample ID	Composition	Sintered Temp	Time	Pressure	Remarks
C1	SS316 (75%) + Al_2O_3, ZrO_2, HA (25%)	900°C	15 min	15 kg/cm²	Clean surface
C2	SS316 (70%) + Al_2O_3, ZrO_2, HA (30%)	900°C	15 min	15 kg/cm²	Slightly brittle
C3	SS316 (70%) + Silicon (30%)	900°C	15 min	15 kg/cm²	Dense and hard sample

Source: Authors (2026).

The applied load for all compactions was fifteen kilogram per square centimeter, ensuring that the green pellets had enough strength to withstand handling and the subsequent thermal treatment. Hardness values were obtained for each of the composite pellets to evaluate the effect of reinforcement on mechanical strength. The data are summarized in Table 6.

Table 6: Hardness results.

Sample ID	Composition	Hardness (HV)	Remarks
C1	SS316 (75%) + Al_2O_3, ZrO_2, HA (25%)	539.3	Moderate hardness
C2	SS316 (70%) + Al_2O_3, ZrO_2, HA (30%)	380.3	Higher ceramic effect
C3	SS316 (70%) + Silicon (30%)	281.5	Possibly highest

Source: Authors (2026).

The pellet with 25% ceramic reinforcement, C1, recorded a hardness of 539.3 HV. This value indicates that the addition of the ceramic blend contributed significantly to resistance against localized deformation. The second pellet, C2, which contained a higher proportion of ceramic reinforcement, showed a hardness of 380.3 HV. Although still higher than unreinforced stainless steel, the reduction compared with C1 suggests that an excessive number of ceramic particles may have influenced the bonding between the steel matrix and the reinforcement, leading to reduced compactness [36] [39]. The third sample, C3, which contained silicon as the reinforcing phase, recorded a hardness of 281.5 HV. This lower value reflects the influence of silicon, which, although beneficial for strength improvements in some alloys, can form brittle phases during sintering that do not contribute positively to indentation resistance. The slightly brittle nature observed in C3 further supports this behaviour.

The comparison of the hardness values shows that the ceramic mixture of alumina, zirconia and hydroxyapatite was more effective in improving the mechanical resistance of the SS matrix compared with silicon. The moderate hardness obtained for C1 suggests an optimal balance between ceramic content and matrix consolidation, whereas the decreased hardness in C2 indicates that increasing ceramic content beyond a certain level may introduce porosity or reduce the bonding efficiency [40]. The silicon reinforced pellet exhibited the lowest hardness among the composites, implying that the structure formed during sintering may not support strong load transfer between the matrix and the reinforcing particles [41]. These results emphasize the importance of reinforcement type and composition in determining the final mechanical properties of powder consolidated materials [34].

IV. DISCUSSION

The results obtained from the Spark Plasma Sintered sample demonstrate the effectiveness of this consolidation technique in producing dense metallic pellets with refined microstructural features. The compact exhibited a smooth surface, limited porosity and clearly defined grain boundaries. Such features indicate efficient bonding between the SS 316 particles and suggest that the material would display strong mechanical behaviour, particularly in hardness and resistance to surface deformation [28]. The controlled thermal cycle of the SPS process appears to have played a major role in achieving these favourable characteristics. In contrast, the pellets prepared through the conventional furnace route displayed small variations in appearance and surface quality. Slight shrinkage, mild oxidation and uneven surface regions were observed.These outcomes are consistent with the nature of furnace sintering, where slower heating rates and the absence of applied pressure limit the extent of densification.

Even so, the pellets retained sufficient structural integrity to proceed with corrosion and mechanical evaluation. Their behaviour under different chemical environments provides valuable insight into the performance of SS 316 when processed using a more traditional approach. The composite samples developed with ceramic and silicon reinforcements displayed noticeably different hardness values, confirming that the type and amount of the reinforcing phase strongly influence the final mechanical response. The combination of alumina, zirconia and hydroxyapatite contributed to higher hardness compared with the silicon-based composite, demonstrating the effectiveness of ceramic additions in improving load bearing capability. The slight brittleness observed in the silicon reinforced material suggests that the formed structure may not support strong stress transfer within the matrix.

The corrosion test results further highlight the importance of processing technique and reinforcement content in determining the chemical stability of the pellets. Differences in mass loss among the samples immersed in distilled water, seawater and sodium hydroxide solutions show that the method of sintering influences surface quality and passive layer formation. The interaction of reinforcement phases with aggressive media will also help identify which composite formulation offers the most favourable balance between mechanical strength and environmental resistance. Taken together, the findings show that SPS provides superior densification and microstructural refinement, ceramic reinforcements enhance hardness and environmental exposure reveals the durability of the produced materials. These observations create a clear connection between processing conditions, microstructure and final performance.

V. CONCLUSIONS

The investigation shows that stainless steel 316 processed through powder metallurgy responds strongly to the choice of sintering technique and reinforcement strategy.

The SPS route delivered the most favourable outcome, producing a compact with very high density, smooth surface finish and a fine microstructure with minimal porosity. Elemental analysis confirmed that the chemical composition remained stable, indicating that the rapid current assisted heating preserved alloy integrity. These results position the SPS method as the most effective route in this study. The furnace sintered samples displayed moderate bonding and predictable mass loss during corrosion testing. Among the exposure conditions, seawater produced the most pronounced degradation due to chloride activity, while distilled water showed minimal effect. The alkaline solution caused surface film changes but lower mass loss compared with the salt environment.

The reinforced composites demonstrated clear variations in hardness. The mixture containing twenty five percent ceramic reinforcement produced the highest hardness value, indicating improved resistance to deformation. The silicon-based composite displayed lower hardness and a slightly brittle character, showing that reinforcement type has a significant impact on mechanical response. Future work should extend the mechanical evaluation to tensile and compression behaviour, explore thermal and electrical properties and optimize sintering conditions for even greater densification. Additional characterization through X ray diffraction, transmission electron microscopy and biological testing for hydroxyapatite-based composites will help establish the full potential of these material systems.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Imthiyas A, Bubesh Kumar D and Sangeetha Krishnamoorthi S.
Methodology: Imthiyas A, Bubesh Kumar D.
Investigation: Imthiyas A, Bubesh Kumar D.
Discussion of results: Imthiyas A, Bubesh Kumar D .
Writing – Original Draft: Imthiyas A.
Writing – Review and Editing: Imthiyas A , Bubesh Kumar D , Sangeetha Krishnamoorthi S , Vasanth kumar J and Dhanush kumar.
Resources: Imthiyas A.
Supervision: Bubesh Kumar D and Sangeetha Krishnamoorthi S.
Approval of the final text: Imthiyas A, Bubesh Kumar D and Sangeetha Krishnamoorthi S, Vasanth kumar J and Dhanush kumar.

VII. REFERENCES'

[1]F. L. Serafini et al., "Materials Characterization Microstructure and mechanical behavior of 316L liquid phase sintered stainless steel with boron addition," Mater. Charact., vol. 152, no. October 2018, pp. 253–264, 2019, doi: 10.1016/j.matchar.2019.04.009.

[2]S. Nitriding et al., "Optimization of Sintering Parameters of 316L Stainless Steel for In-Situ Nitrogen Absorption and," pp. 1–20, 2020.

[3]S. Ali, A. Majdi, A. Rani, Z. Baig, S. W. Ahmed, and G. Hussain, "Biocompatibility and corrosion resistance of metallic biomaterials," vol. 38, no. 5, pp. 381–402, 2020.

[4]M. Kaur and K. Singh, "Review on titanium and titanium-based alloys as biomaterials for orthopaedic applications," Mater. Sci. Eng. C, vol. 102, no. February, pp. 844–862, 2019, doi: https://doi.org/10.1016/j.msec.2019.04.064.

[5]A. Moridi, "Biomedical Applications of Metal Additive Manufacturing: Current State-of-the-Art and Future Perspective," Am. J. Biomed. Sci. Res., vol. 7, no. 1, pp. 6–10, 2020, doi: 10.34297/ajbsr.2020.07.001103.

[6]J. Ni et al., "Three-dimensional printing of metals for biomedical applications," Mater. Today Bio, vol. 3, no. August, 2019, doi: 10.1016/j.mtbio.2019.100024.

[7]A. B. Kale, B. Kim, D. Kim, E. G. Castle, M. Reece, and S. Choi, "Materials Characterization An investigation of the corrosion behavior of 316L stainless steel fabricated by SLM and SPS techniques," Mater. Charact., vol. 163, no. February, p. 110204, 2020, doi: 10.1016/j.matchar.2020.110204.

[8]E. Hug, C. Folton, J. Papin, K. Tabalaiev, and G. Marnier, "Ultrafine Grain 316L Stainless Steel Manufactured by Ball Milling and Spark Plasma Sintering : Consequences on the," 2024.

[9]Z. Zhan and H. Li, "Machine learning based fatigue life prediction with effects of additive manufacturing process parameters for printed SS 316L," Int. J. Fatigue, vol. 142, no. July 2020, p. 105941, 2021, doi: 10.1016/j.ijfatigue.2020.105941.

[10]M. Al-Amin et al., "Analysis of hybrid HA/CNT suspended-EDM process and multiple-objectives optimization to improve machining responses of 316L steel," J. Mater. Res. Technol., vol. 15, pp. 2557–2574, 2021, doi: 10.1016/j.jmrt.2021.09.074.

[11]H. Tetsuka and S. R. Shin, "Materials and technical innovations in 3D printing in biomedical applications," J. Mater. Chem. B, vol. 8, no. 15, pp. 2930–2950, 2020, doi: 10.1039/d0tb00034e.

[12]L. F. Velásquez-García and Y. Kornbluth, "Biomedical Applications of Metal 3D Printing," Annu. Rev. Biomed. Eng., vol. 23, pp. 307–338, 2021, doi: 10.1146/annurev-bioeng-082020-032402.

[13]W. Wang et al., "3D printing of PLA/n-HA composite scaffolds with customized mechanical properties and biological functions for bone tissue engineering," Compos. Part B Eng., vol. 224, no. July, p. 109192, 2021, doi: 10.1016/j.compositesb.2021.109192.

[14]S. M. Yusuf and N. Gao, "Influence of energy density on metallurgy and properties in metal additive manufacturing," Mater. Sci. Technol. (United Kingdom), vol. 33, no. 11, pp. 1269–1289, 2017, doi: 10.1080/02670836.2017.1289444.

[15]G. Iacob, V. G. Ghica, M. Buzatu, T. Buzatu, and M. Ionut, "Composites : Part B Studies on wear rate and micro-hardness of the Al / Al 2 O 3 / Gr hybrid composites produced via powder metallurgy," vol. 69, pp. 603–611, 2015, doi: 10.1016/j.compositesb.2014.07.008.

[16]H. Li, M. Ramezani, M. Li, C. Ma, and J. Wang, "Tribological performance of selective laser melted 316L stainless steel," Tribol. Int., vol. 128, no. April, pp. 121–129, 2018, doi: 10.1016/j.triboint.2018.07.021.

[17]V. Vishnu, T. R. Prabhu, M. Imam, and K. P. Vineesh, “Experimental Investigation into the Dry Reciprocating Wear Behavior of Additively Manufactured Austenitic Stainless Steel (316L) Alloys,” Jom, vol. 76, no. 6, pp. 3081–3098, 2024, doi: 10.1007/s11837-024-06491-9.

[18]S. N. Alam and L. Kumar, “Mechanical properties of aluminium based metal matrix composites reinforced with graphite nanoplatelets,” Mater. Sci. Eng. A, vol. 667, pp. 16–32, 2016, doi: 10.1016/j.msea.2016.04.054.

[19]P. Samal, P. R. Vundavilli, A. Meher, and M. M. Mahapatra, “Recent progress in aluminum metal matrix composites: A review on processing, mechanical and wear properties,” J. Manuf. Process., vol. 59, no. September, pp. 131–152, 2020, doi: 10.1016/j.jmapro.2020.09.010.

[20]E. Hug and C. Keller, “Passivation properties of ultrafine grain 316L stainless steels manufactured by ball milling and spark plasma sintering,” Mater. Chem. Phys., vol. 309, no. August, p. 128398, 2023, doi: 10.1016/j.matchemphys.2023.128398.

[21]K. Ravi Kumar, T. Pridhar, and V. S. Sree Balaji, “Mechanical properties and characterization of zirconium oxide (ZrO2) and coconut shell ash(CSA) reinforced aluminium (Al 6082) matrix hybrid composite,” J. Alloys Compd., vol. 765, pp. 171–179, 2018, doi: 10.1016/j.jallcom.2018.06.177.

[22]L. Kang, F. Chen, B. Wu, X. Liu, and H. Ge, “Mechanical properties and microstructure of laser-cladding additively manufactured 316L stainless steel sheets,” J. Constr. Steel Res., vol. 199, no. December 2021, p. 107603, 2022, doi: 10.1016/j.jcsr.2022.107603.

[23]M. Zhao, H. Wu, J. Lu, G. Sun, and L. Du, “Materials Characterization Effect of grain size on mechanical property and corrosion behavior of a metastable austenitic stainless steel,” Mater. Charact., vol. 194, no. September, p. 112360, 2022, doi: 10.1016/j.matchar.2022.112360.

[24]S. Rangrej, S. Pandya, and J. Menghani, “Effects of reinforcement additions on properties of aluminium matrix composites - A review,” Mater. Today Proc., vol. 44, pp. 637–641, 2021, doi: 10.1016/j.matpr.2020.10.604.

[25]F. Chen et al., “Tribological behavior and mechanism of h-BN modified copper metal matrix composites paired with C/C–SiC,” Tribol. Int., vol. 153, no. July 2020, 2021, doi: 10.1016/j.triboint.2020.106561.

[26]Y. C. Wu et al., “The superior properties of spark plasma sintered La-Fe-Si magnetocaloric alloys,” Mater. Res. Bull., vol. 155, no. July, p. 111974, 2022, doi: 10.1016/j.materresbull.2022.111974.

[27]Z. Hu, Z. Zhang, X. Cheng, F. Wang, Y. Zhang, and S. Li, “A review of multi-physical fi elds induced phenomena and effects in spark plasma sintering : Fundamentals and applications,” vol. 191, 2020.

[28]S. Ali et al., “Microstructure and Mechanical Properties of Modified 316L Stainless Steel Alloy for Biomedical Applications Using Powder Metallurgy,” 2022.

[29]C. K. F. Fang, “State of the art of bioimplants manufacturing : part II,” Adv. Manuf., vol. 6, no. 2, pp. 137–154, 2018, doi: 10.1007/s40436-018-0218-9.

[30]A. Azeez et al., “A Review of Additive Mixed-Electric Discharge Machining : Current Status and Future Perspectives for Surface Modification of Biomedical Implants,” vol. 2017, 2017, doi: 10.1155/2017/8723239.

[31]C. K. F. Fang, “State of the art of bioimplants manufacturing : part I,” Adv. Manuf., vol. 6, no. 1, pp. 20–40, 2018, doi: 10.1007/s40436-017-0207-4.

[32]E. Mardaras and I. Fern, “Influence of Severe Surface Plastic Deformation Induced by Shot Peening on Microstructure and Corrosion Resistance of fine grained 316 L stainless steel,” vol. 231, no. January, 2024, doi: 10.1016/j.corsci.2024.111988.

[33]X. Yue, Z. Yang, L. Huang, L. Zhang, J. Li, and Z. Xue, “Journal of Materials Science & Technology Passivation characteristics of ultra-thin 316L foil in NaCl solutions,” J. Mater. Sci. Technol., vol. 127, pp. 192–205, 2022, doi: 10.1016/j.jmst.2022.01.043.

[34]S. Robi, X. Wang, M. Brochu, and S. Omanovic, “Effects of crystallographic orientation on the corrosion behavior of stainless steel 316L manufactured by laser powder bed fusion,” Corros. Sci., vol. 196, no. September 2021, p. 110009, 2022, doi: 10.1016/j.corsci.2021.110009.

[35]A. Racot, I. Aubert, M. Touzet, and M. Demesy, “Statistical analysis of the pitting corrosion induced by potentiostatic pulse tests of wrought and SLM 316L stainless steels,” vol. 197, no. October 2020, 2022, doi: 10.1016/j.corsci.2021.110036.

[36]J. Choi, G. Lee, J. Song, W. Lee, and J. Lee, “Sintering behavior of 316L stainless steel micro – nanopowder compact fabricated by powder injection molding,” Powder Technol., vol. 279, pp. 196–202, 2015, doi: 10.1016/j.powtec.2015.04.014.

[37]E. Liu, “E ff ect of strain-induced martensite on the tribocorrosion of AISI 316L austenitic stainless steel in seawater,” pp. 44923–44932, 2017, doi: 10.1039/c7ra07318f.

[38]Q. He et al., “Microstructure and properties of 316L stainless steel foils for pressure sensor of pressurized water reactor,” Nucl. Eng. Technol., vol. 53, no. 1, pp. 172–177, 2021, doi: 10.1016/j.net.2020.06.006.

[39]A. Parveen, N. R. Chauhan, and M. Suhaib, “Mechanical and tribological behaviour of Al-ZrO2 composites: A review,” Lect. Notes Mech. Eng., pp. 217–229, 2019, doi: 10.1007/978-981-13-6469-3_20.

[40]E. E. Feldshtein, L. N. Dyachkova, and J. Patalas-Maliszewska, “On investigating the microstructural, mechanical, and tribological properties of hybrid feGr1/Sic/Gr metal matrix composites,” Materials (Basel)., vol. 14, no. 1, pp. 1–12, 2021, doi: 10.3390/ma14010174.

[41]S. Sahoo, S. Samal, and B. Bhoi, “Fabrication and characterization of novel Al-SiC-hBN self-lubricating hybrid composites,” Mater. Today Commun., vol. 25, no. March, p. 101402, 2020, doi: 10.1016/j.mtcomm.2020.101402.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 428-447. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3030

RESEARCH ARTICLE OPEN ACCESS

AI APPLICATIONS FOR DRONE OF THINGS IN THE POWER SECTOR

Manjunatha H. M.[*1], Madan H. T.[2], Poshitha B.[3], Swapna H.[4], Aruna B.[5], Pramodh H. K.[6] and Lokesh Mallaiah[7]

[1]Electrical and Electronics Engineering, Bapuji Institute of Engineering and Technology, Davanagere-577004, Karnataka, India.
[2]NITTE (Deemed to be University), Department of Electronics and Communication (Advanced Communication Technology), NMAM Institute of Technology, Nitte, Karkala Taluk, Udupi, 574110, Karnataka, India
[3]Electrical and Electronics Engineering, Adichunchanagiri Institute of Technology- 577102, Chikkamagaluru -577004, Karnataka, India.
[4]Electrical and Electronics Engineering, ATME College of Engineering, Mysuru - 570028, Karnataka, India.
[5]Electronics and Communication Engineering ,BGS institute of Technology, Adichunchanagiri University, BG Nagar-571448, Karnataka, India.
[6]Electrical and Electronics Engineering, Malnad college of Engineering, Hassan - 573202, Karnataka, India.
[7]Electrical and Electronics Engineering, Seshadripuram Institute of Technology, Mysuru - 571311, Karnataka, India.

[1]https://orcid.org/0000-0001-6651-7131, [2]http://orcid.org/0000-0002-3126-9737, [3]http://orcid.org/0009-0007-0893-4325
[4]http://orcid.org/0000-0001-9086-5507, [5]http://orcid.org/0009-0006-9530-5131, [6]http://orcid.org/0009-0003-0833-7644
[7]http://orcid.org/0000-0003-4873-4719

Email: *manjunath.hm1986@bietdvg.edu, madan.ht@nitte.edu.in, drposhitha@gmail.com, swapnah90@gmail.com, baruna@bgsit.ac.in, pramodh.hk@gmail.com, drlokeshmallaiah@gmail.com

ARTICLE INFO

Article History
Received: December 1, 2025
Revised: December 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Artificial Intelligence,
Drone of Things,
Power Sector,
Internet of Things,
Infrastructure Inspection,
Predictive Maintenance.

ABSTRACT

Drones that are equipped with artificial intelligence and interconnected with the Internet of Things have revolutionized the monitoring, maintaining and managing of power infrastructure over the last decade. This Paper provides an overview of the background of the Drone of Things and basic components-and sets the stage for exposure that is much more detailed than what has been discussed on Artificial Intelligence applications in the electric power sector up to this point. Examining the cases highlights the significance of Artificial Intelligence enabled Drone of Things systems, particularly in enhancing infrastructure inspection, predictive maintenance, disaster response, energy theft detection and planning sites for renewable energy. This paper discusses fundamental Artificial Intelligence approaches utilized in practical applications such as computer vision, machine learning, deep learning, and natural language processing. This paper also addresses possible challenges to their implementation-such as regulation, integration, and the accuracy of Artificial Intelligence algorithms-and the limitations of Artificial Intelligence -enabled Drone of Things systems. In the future, new developments in 5G, edge computing, swarm intelligent systems, and advanced sensing systems will provide a wide range of new ways that Artificial Intelligence powered Drone of Things systems can be utilized in the electric sector. The following analysis gives an in-depth look into how Artificial Intelligence enabled Drone of Things are important for operation optimization, reliability enhancement, and providing safety to power systems. Those technologies will lead to many advances and improvements in the management and distribution of electricity in the energy sector over the coming years.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

A power sector in profound and exciting change with the advent of convergence of drones, Internet of Things (IoT) and Artificial Intelligence (AI). This powerful synergy is giving life to the Drone of Things (DoT), or rather, the network of intelligent and interconnected drones that revolutionize how we manage, maintain and optimize power infrastructure [1].

Imagine drones mapping a continent. They fly meters and miles to scan transmission lines when all the signs of distress and fatigue can be detected consciously. They can detect hazards such as vegetation intrusion and even predict possible equipment failure, sometimes before it happens. This is the power of AI-powered DoT (AI-DOT) in the power sector [2], [3].

I.1 THE CONVERGENCE OF DRONES, IOT AND AI

It combines drones, IoT and AI technologies to create a complete power infrastructure management system [4], [5]. Each of these technologies plays a key part in building an all-in-one and intelligent system to manage power infrastructure. Drones have grown from basic remote-controlled gadgets to advanced machines that gather data. High-resolution cameras, thermal imaging sensors, light detection and ranging (LiDAR) technology and even corona discharge detectors are among their capabilities, making them indispensable for monitoring the power grid. They can reach inaccessible and dangerous spots such as high-voltage transmission towers, dense forests and offshore wind farms, thus saving people from risking themselves.

They can obtain an accurate and detailed perspective of the power assets through visual data that is not feasible to gather from the ground [5]. The backbone of this smart system is the IoT, which connects a significant network of devices and sensors that can be embedded in various places in the power grid. The network offers real-time information concerning constant performance, environmental conditions, as well as hazards within the system. Secondary substations might contain temperature, vibration or oil level sensors etc. In the event of their weather stations, they can have wind speed, temperature, and humidity level measurements. This information with the visual information gathered by the drones makes an exhaustive digital twin of the physical grid.

An integrated view of its health and performance is offered. AI has a major influence on this networked system. Complex formulas sift through massive amounts of information from drones and IoT devices spotting patterns, differences, and oddities that might signal upcoming problems. In other words, AI can detect small shifts in a transformer's heat pattern that could mean it's about to fail, or even a slight curve in a power line tower that could lead to disaster. This allows for fixing things before they break and planning repairs in advance, which stops expensive power outages and ensures people get electricity they can count on [6].

I.2 THE DRONE OF THINGS (DOT) IN THE POWER SECTOR

The merging of drones, IoT, and AI leads to the idea of the DoT, a system of smart drones working together. This has a big impact on how power systems work. Drones aren't just remote-controlled devices anymore. They're now smart helpers that can choose what to gather and what AI programs should look at.They can adapt their flight paths in real time to focus on areas of concern, communicate with other drones to coordinate inspections or even return to base for recharging or data transfer without human intervention [1-6]. In the power sector, DoT is transforming various aspects of operations, including:

- DoT offers an autopilot inspection of power lines, towers, substations and any infrastructure required. These drones will be able to sense symptoms of corrosion, damage and vegetation encroachment, among other hazards useful for maintenance planning and assessment of potential risks.
- The AI powered DoT, on the distribution equipment will predictively maintain the devices because it analyzes data from drones and sensors IoT to identify possible equipment failures that may occur. A proactive approach reduces downtime, lowers repair costs and enhances reliability of power supply.
- In case of any power cut or natural calamity, DoT can be immediately deployed to assess the damage and identify the cause of the problem and guide the repair crews to the affected areas. The rapid capability of this deployment reduces downtime and allows for quicker power restoration.
- DoT can be used for security surveillance of power infrastructure to detect vandalism, access by unauthorized people and other security threats.

I.3 ADVANTAGES OF AI-POWERED DOT IN POWER SYSTEMS

The adoption of AI-powered DoT in the power sector offers numerous advantages:

- **Improved Efficiency and Productivity:** DoT automates tasks that were earlier time-consuming and labour-intensive, including inspecting power lines and equipment. This significantly improves efficiency and productivity by reducing man-hours and inspection time, hence generating cost savings.
- **Improved Safety:** It ensures that droids can reach dangerous or inaccessible regions, thus minimizing risks to human workers. It enhances safety and reduces chances of accidents.
- **Accurate and Reliable Systems:** AI algorithms can analyze data much more accurately than humans, with more accurate detection of faults, thus much higher reliability for power systems. The chances of a power outage become much lower and therefore maintain greater supply stability.
- **Data Insights:** Artificial intelligence will be able to draw insightful inferences from the data and hence enable proactive maintenance, optimal utilization of resources, better planning of things, which ultimately will manifest as better decisions and thus improved efficiency in operational activities.
- **Reduce Downtime and Outages:** Predictive maintenance and rapid response capabilities of AI-powered drones minimize downtime, thereby averting costly power outages. This is also instrumental in building reliability into the power supply while reducing the economic impact of disruptions.
- **Environmental Benefits:** DoT can therefore help in the environmental sustainability through the following ways: Maximizing the generation of renewable energy and reducing reliance on fossil fuels. For example, drones inspect solar panels for any damage or soiling that helps maximize optimal energy production.

The confluence of drones, IoT, and AI opens an era for intelligent management and maintenance of power infrastructure. The DoT or Drone of Things is a promise that stands poised to yield improvements in efficiency, safety, accuracy, and data-driven decisions in the areas of power sectors. Based on this technology, the power sector will finally be able to optimize operations, reduce costs, and improve reliability to help build a more sustainable future.

II. AI-DRIVEN DRONE APPLICATIONS ACROSS THE POWER VALUE CHAIN

AI-DoT are transforming the power sector by enabling efficient and intelligent automation across the entire value chain, from generation to distribution.

II.1 GENERATING PLANTS

II.1.1 Solar Farm Inspection

AI-DoT are revolutionizing solar farm inspections by providing rapid, accurate and comprehensive assessments of vast arrays of solar panels. Let's delve into the technical aspects of how these drones are enhancing efficiency and optimizing energy production. Fig.1 shows the block diagram of overall AI-DoT system for Solar Farm Inspection [7], [8].

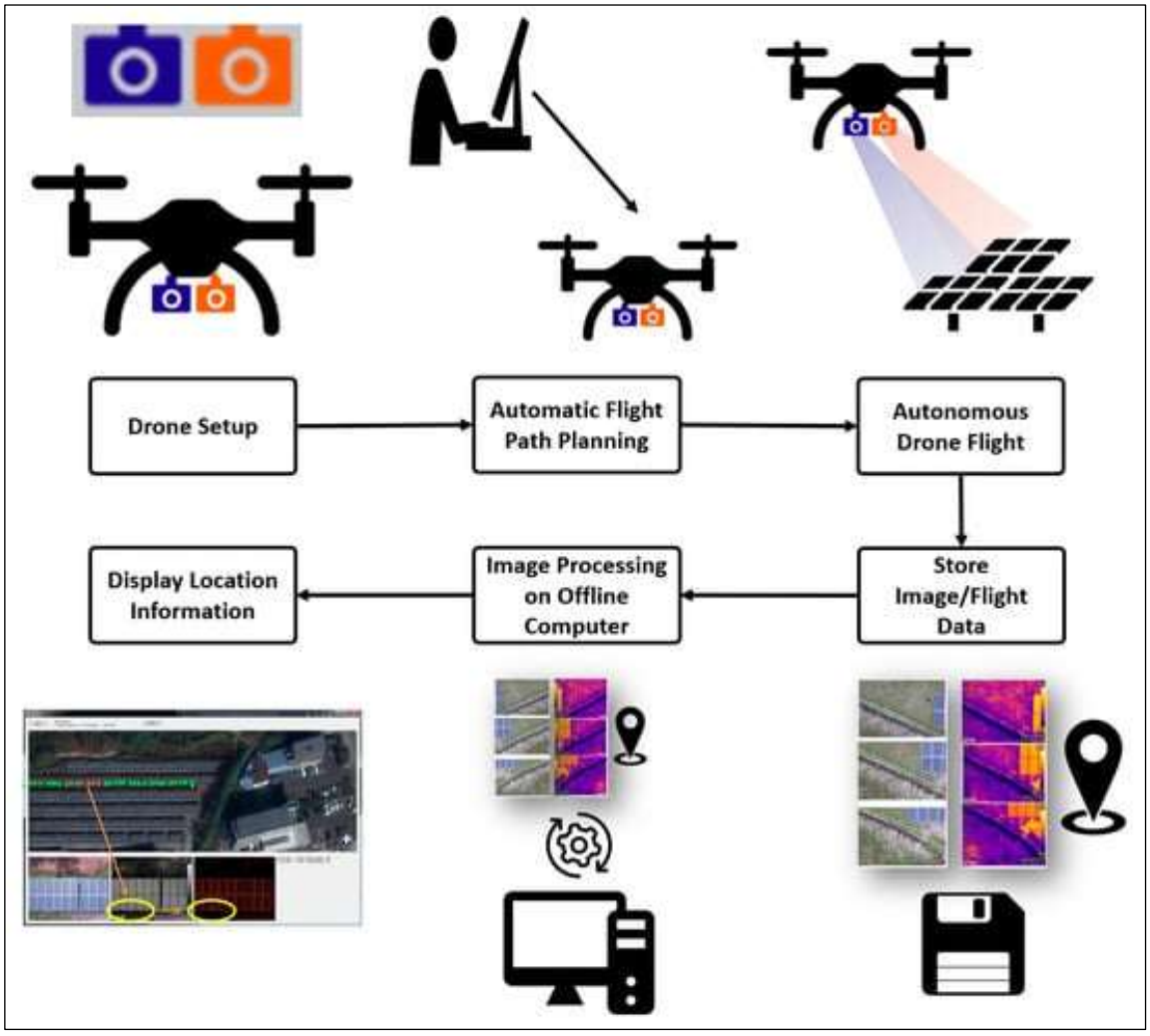


Figure 1: Block diagram of overall AI-DoT system for Solar Farm Inspection.
Source: [9].

II.1.1.1 Panel Defect Detection

The advent of high-resolution imaging owing to drone technology has changed the whole narrative of maintaining solar energy plants. The drones come equipped with high power cameras enabling them take clear images of the solar panels even at an individual level. The pictures are then processed by complex AI systems with most relying of Convolutional Neural Network (CNN) technologies to inspect the images as though a human inspector was being used. These systems are capable of identifying numerous defects that can cause poor performance of a solar panel without any human interaction. Like an AI detective going through each panel with a clear lens, it can detect minute cracks that can cause failure down the line or internal separation of the panels(delamination). The AI will also detect thermal hot spots on the solar panels, which are pronounced areas with high temperatures that either indicate electrical faults or cell damage and can excessively lower energy production or even cause a fire.

Loses due to these structural elements are not the only obstacles that this AI has outsmarted as it has the ability to measure as well the soiling on the panels worn towards the hostile external environment hardening process making energy generation difficult. This automated analysis is not just looking for an issue, but rather determining how best to deal with it. The AI produces a comprehensive report on each damage with its exact location and severity allowing the maintenance team attend repairs and schedule clean ups accordingly. Principle of focusing on most damaging issues is adopted so that costs and time take in implementing maintenance measures is reduced. In the long run, this ensures harnessing more energy, minimizing stand-by time of the solar farms and facilitating better management of the solar farms. Fig.2 shows the Image of overall PV power plant followed by automatically planned flight path and location of defective PV modules respectively [9].

II.1.1.2 Soiling Analysis

Picture drones that have been designed to accommodate advanced cameras which can see beyond the capacity of the human eye. This is the potential of multi-spectral and hyper-spectral imaging which is changing knows and manages soiling of solar panels. These technologies enable the use of drones to obtain images in several light wavelengths revealing the composition of the soiling that would have remained concealed. Multispectral cameras take a picture only in a handful of selective bands of the electromagnetic spectrum like the red, green and blue visual bands with extra bands outside that are seen in. With this characteristic, they are able to differentiate between some basic types of soiling such as dust and plant materials. Hyperspectral cameras, in comparison, take pictures in hundreds of wavelength bands to give a clearer view of the soiling than a 'seeing' camera. Thus, enabling specific things to be determined such as types of dust, pollen, or even faces from birds.

AI analysis tools do the rest and provide details on the soiling level and soiling type of every individual panel. It's like having a dirt analysis of every solar panel in your system. Such information is critical in the effective management of cleaning operations. The areas that get affected the most will be cleaned eliminating wastage of time and resources. In addition, the evaluation is useful in looking for the right cleaning solution in as much as different types of soiling may call for different cleaning approaches. For example loosening dried grime of bird stools will require a different technique from dry wing surfaces. Through this targeted style of cleaning, solar farm owners can appreciate the necessity of the cleaning exercise and as such, keep the cleaning practice productive and the energy output high while curbing energy losses that arise from excessive cleaning[10], [11].

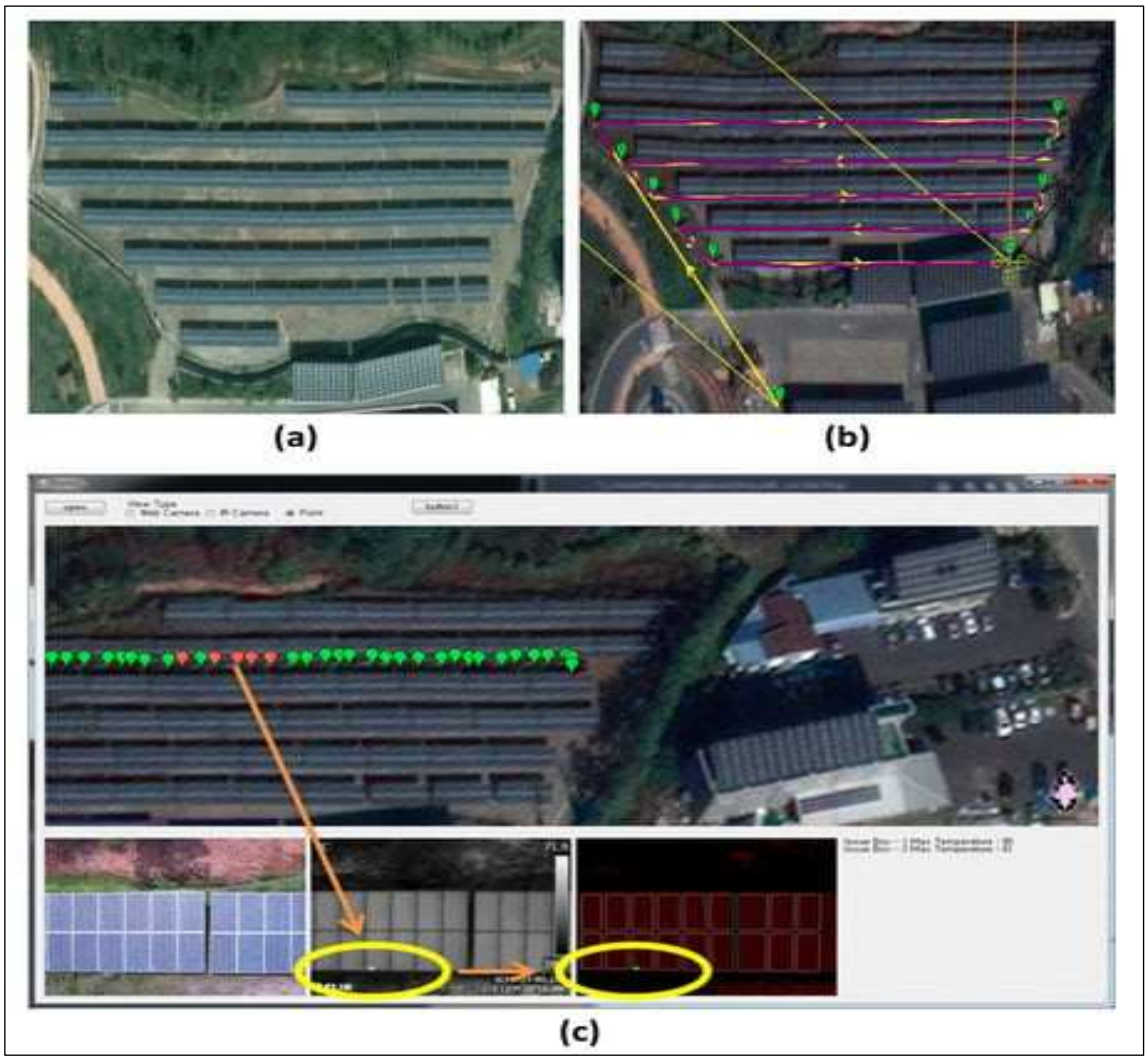


Figure 2: (a) Image of overall PV power plant; (b) image showing the automatically planned flight path; and (c) the final result showing the location of defective PV modules.
Source: [10].

II.1.1.3 Vegetation Management

Drones equipped with LiDAR sensors and high-resolution cameras are revolutionizing the management of solar farms by literally adding a new perspective to it. These technologies facilitate the production of realistic 3D models of the solar farm and its immediate environment, which assists in the understanding of vegetation management in great detail. LiDAR entails the transmission of laser beams and measuring how long it takes for those beams to reflect back from any object to form a 'point cloud' of the surface represented. Thus, this data when incorporated with an image of high resolution it becomes possible to mould the three-dimensional model of the solar farm's profile, including its landscape, overgrowth and solar panels[12], [13]. Analyzing such models for, among other things, vigorous growth around the solar array ie where trees, bushes or other plants are growing up to the edge of the solar panels is performed by AI.

Each leaf, even that of a small shrub shading a few millimeters of the panel can reduce its energy production significantly, or in some cases destroy entirely the panel. Thanks to this feature, the AI also suggests which areas are the most important for receiving attention on the suppression of foliage growth. This is a more effective way of managing vegetation as most of it is directed towards the target only. Vegetation management is simplified as maintenance crews now only go out to cut back or remove offending shading vegetation. In addition, this helps in capitalizing on solar energy within the solar farm in a more effective way by addressing potential shading issues that might arise in future and ensuring renewed efficiency of the solar farm. Fig.3 shows the LiDAR point cloud colored by Sitemark followed by Visual orthomosaic and contour lines based on the colored LiDAR point cloud [14], [15].



(a) **(b)**

Figure 3: (a) LiDAR point cloud colored by Sitemark (b)Visual orthomosaic and contour lines based on the colored LiDAR point cloud.
Source: [15].

II.1.2 Wind Turbine Inspection

AI-powered drones are transforming wind turbine inspections, enabling safer, faster and more accurate assessments of these towering structures. Let's delve into the technical aspects of how drones are enhancing the efficiency and safety of wind turbine maintenance [16]. Fig.4 shows the Wind turbine drone inspection system and Fig.5 summarise the flowchart of the automated wind turbine damage suggestion system.



Figure 4: Wind turbine drone inspection system.
Source: Authors (2026).

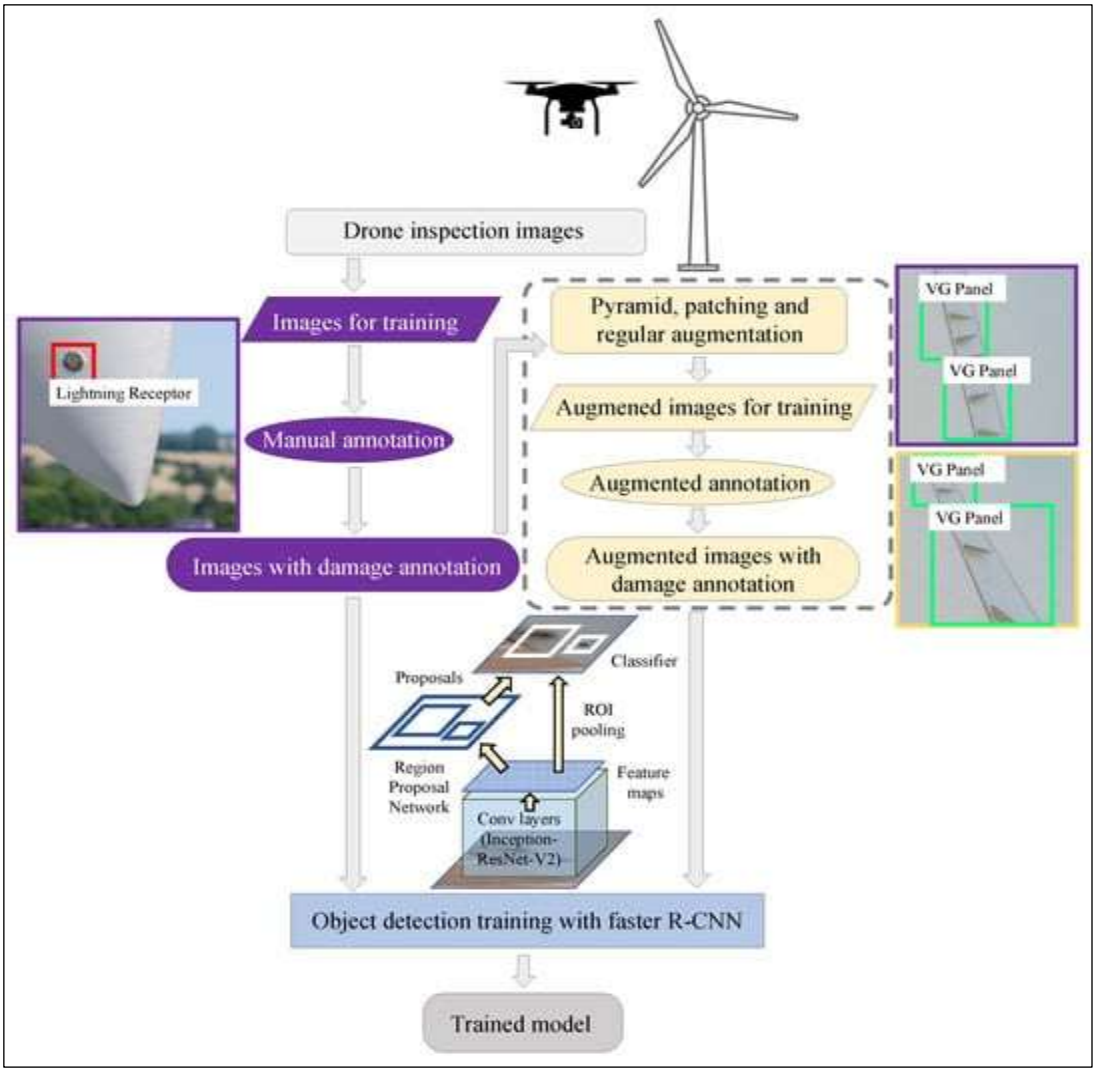


Figure 5: Flowchart of the automated wind turbine damage suggestion system.
Source: [17].

II.1.2.1 Blade Damage Assessment

The ability to obtain high resolution images through drones is altering the process of wind turbine blades inspections. These drones are fitted with sophisticated cameras capable of imaging the entire blade surface, including the tip, the leading and trailing edges. This form of imaging then undergoes high level processing in visual computing packages including the semi-automated Convolutional Neural Networks that play a role of a blade inspector. Such systems of artificial intelligence are proficient in damage assessment. They are able to spot defects such as cracks which if not treated may widen resulting to a breakage of a blade. They are also able to assess the level of blade wear at distinct areas such as the leading edge where blades face erosion due to rain, hail and dust. In addition, the AI system can also assess the extent of other damages such as burn marks, delamination and structural damages that are caused due to lightning striking the turbine.

Even more specialized algorithms deal with the leading edge of blades, the area most prone to erosion, in order to assess the type and the amount of damage inflicted is erosion. Blades can also be analyzed by Drones equipped with LiDAR sensors that can build 3D models of the blades. This allows a better understanding of the geometric construction of the blades, which improves damage evaluation. The AI then architectural powers reports after processing the images and the 3D model whereby the report includes the exact location and the level of seriousness of each defect. This way they are able to guide the maintenance teams on which repairs need to be done first, thereby avoiding the low-cost repairs that may lead to failures and addressing the most important issues to extend the life of the wind turbine [17].

II.1.2.2 Gearbox and Generator Monitoring

Consider a drone as a flying doctor who has a medical toolkit, in this case, listening to the heartbeat of the wind turbine. What this means in laying terms is advanced sensor integration and AI-based monitoring gearbox and generator health. Today, drones are also fitted with sensors that are able to take the status of these major elements and gather lots of information about their condition. To this end, there are vibration sensors, which are placed in the gearbox and the generator to measure the shakes within the apparatus like the sound of a stethoscope. These shakes are likely to change due to wear and tear, misalignment, or other issues. Temperature sensors, which are fitted in sensitive areas, such as the bearings and gears, are analogous to thermometers.

For instance, a bearing may be worn out or run dry and causing excess heat. Last but not least, the use of acoustic sensors is similar to the use of very keen-sounding microphones to listen into the inner workings of the sounds captured. An abnormal sound could signify the presence of a loose component or a chipped gear within the gearbox or the generator. When this information is processed, that is when the true enchantment occurs, as the sensor data gets analysed with the help of AI algorithms in real time. The AI assumes the role of the turbine operator, continuously monitoring the condition of the equipment. It does this by looking for abnormalities, or departures from the norm – excessive vibrations, elevated temperatures, etc.

Additionally, and in the case of machines such as i.e. wind turbines, which have huge amounts of history and utilization, the AI can forecast failure scenarios and adjust repair times accordingly based on previous machine learning. This means that it becomes possible to replace systems without waiting for a breakdown to avoid losses due to downtime and to increase the service life of the turbine. This strategy aiming to minimize down time in turbines maintenance using sensor data and information from AI's calculations is cutting across and changing the corporate marketing of wind turbine maintenance [17].

II.1.2.3 Lightning Protection System Checks

Its impressive capital assets demand that the integrity of the lightning protection system of a wind turbine be upheld. More recently, drones have been incorporated in the safety assessments increasing the range, depth and efficiency of inspection of the lightning protection system. Visualize the drone as a small flying electrician, inspecting every possible part of the system with intensive focus. The drone's high-resolution camera is capable of looking at the lightning receptors, the down conductors connecting the lightning to the ground currents and the ground itself. The drone's inspective abilities extend to advanced observations such as signs of damage, corrosion and loose fittings that can potentially render the system ineffective. However, the inspection does not end up there. Additional significance is also added to the drone by using thermal cameras which are capable of measuring the temperatures of different regions and thus identifying the presence of hot spots within the lightning protection system. Hot spots are variations in temperature that are areas of trouble within the system.

Hot spots are not always visible, yet they are the most common suspects that usually cause the system from properly dissipating the electricity into the ground [17]. In order to improve the inspection further, a special sensor installed on the drone can be used to check electrical values of the grounding system. It's like examining the power lines to see whether they are efficient and in a good condition. It can be estimated that if the reading was found to be high at this point, the high resistance would suggest the existence of a fault in the grounding system which could lead to damage to the turbine if struck by lightning. Finally, all this visual, thermal and electrical data is processed by AI. The AI behaves like a doctor who studies the health state of the lightning protection system and looks for possible reasons of its malfunction. This comprehensive and automated approach to inspection ensures that the lightning protection system is functioning optimally, providing reliable protection for the wind turbine and maximizing its operational lifespan.

II.1.3 Hydroelectric Dam Inspection

Drones with AI seem to be real helpers when it comes to checking hydroelectric dams and doing a wide range of quick and thorough assessments of important infrastructure. Let's get into the details of how drones improve safety and performance at hydroelectric dams.

II.1.3.1 Concrete Crack Detection

This is where concrete structures such as bridges, dams, buildings etc., require proper inspection and maintenance to make it safe and long-lasting. Traditionally, this was a laborious manual visual inspection that takes up considerable time with a high possibility of human error. However, the development in AI-based DoT is transforming this into a safer and more efficient and accurate technique of concrete crack detection. The high-resolution cameras in the drones capture images of the surface of concrete from various angles, and thus provide a panoramic view of the structure. This enables them to cover extensive coverage quickly, thus saving much inspection time. These images are then analyzed for cracks by sophisticated computer vision algorithms often based on CNNs. AI algorithms can be trained on massive images with cracks labeled so that there will be identification and classification of the cracks based on size, shape and direction. Accuracy is also ensured by arming the drones with LiDAR sensors to create 3D models of the structure.

The surface representation of the detailed crack detection would be much more precise, especially in complex or hard-to-reach areas. The 3D models and images will then be analyzed by the AI algorithms that automatically identify the crack existence and generate reports with considerable detail about them. The reports contain precise location information and assessments of the severity, which would help engineers prioritize the necessary repair and maintenance work. This technology has various advantages. It ensures safety because it can reach areas which are inaccessible by humans by opening it up for the drones to inspect; thereby, reducing the necessity of inspection at a height or confined space. It enhances efficiency through automated inspection that tends to be far faster than the traditional methods. Accuracy is also enhanced since AI algorithms can decipher cracks almost invisible to the naked eye. For example, it provides factual opinions rather than judgments by humans. It may also save time, during the inspection process and be much accurate.

AI-based DoT for concrete crack detection has vast applications within the industrial sector. This can be applied in both bridge inspections, dams, and tunnels to detect weaknesses in the structure before they eventually cause failure. It can be applied for the inspection of buildings ensuring safety and sustainability in high-rise buildings, historical buildings, and other buildings. It can be applied for construction monitoring purposes to identify issues related to concrete beginning from the construction phase even post-construction. AI-powered DoT is changing the way we find cracks in concrete. It's making it safer, faster, and more accurate to check on and keep up with important infrastructure. It is an important technology for keeping concrete structures safe and making sure they last a long time. Its goal is to make the built environment better, stronger, and more sustainable. There are a lot of good things about this technology. Drones can cover large areas quickly and effectively, which is much faster than traditional methods.

It is also accurate because the AI algorithms can find small signs of leaks that people might not see. The repair is done faster, which protects against more damage and saves money in the long run. The most important thing is that the tool saves a lot of water because it can find leaks early and stop them from losing more water. This serves as excellent environmental use since it reduces water wastage as well as protects the surrounding ecosystems from damage brought about by leakages. AI-powered DoT for water leakage identification has several applications. The tool can be used for pipeline inspections, among others such as water distribution networks, oil and gas pipelines or sewage systems. It can be used for leak detection in the dams and reservoirs so that possible damages are avoided along with saving the neighboring communities from destruction. In agriculture, it will detect leaks in irrigation canals so that water will not be wasted and efficiency increases. And generally, it will use in monitoring water bodies so that areas of water loss could be determined to manage our precious water resources well. Fig.6 shows the AI powered DoT for concrete crack detection system.

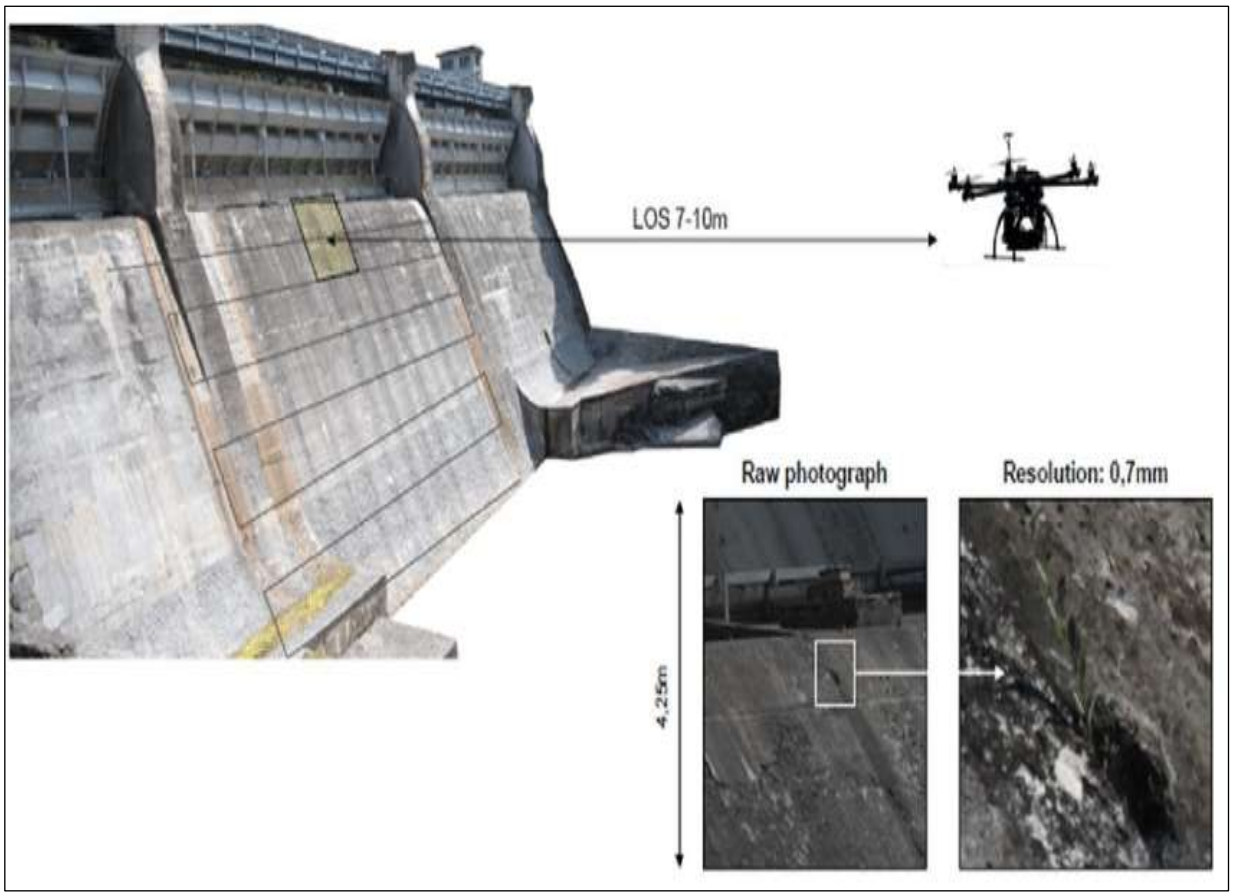


Figure 6: AI Powered DoT for concrete crack detection system.
Source: [18].

II.1.3.2 Sedimentation Monitoring

It has presented one of the major challenges in the management of water resources and hydroelectric power generation due to the buildup of sediment within reservoirs and the waterways. The accumulation decreases the capacity of a reservoir, degrades water quality and in extreme cases can stop the flow in hydropower turbines, thus reducing their efficiency. Initially, manual surveys and sampling of water were the core monitoring activities on sedimentation, which were very costly and time-consuming and offered very little information. However, a Drone of Things powered by AI would revolutionize the whole process much more efficiently, much more accurately, and far more comprehensively. Just imagine drones carrying sonar or LiDAR like underwater surveyors. They fly over the water and send sound waves or laser pulses down through the surface to map the underwater topography. This produces an incredibly detailed 3D map of the reservoir bed showing depth and distribution of sediment. Such drones also carry multispectral cameras capturing images of the water in different wavelengths of light.

This would help analyse the quality of the water and enable the team to identify areas with high sediment concentration, almost like taking an X-ray of water. The real power of this technology lies in the AI algorithms that analyse the data collected by these drones. These can suitably measure the volume of sediment in a reservoir, explain how it is spread, and then continue with fluctuations in levels of sedimentation over a given period. All this makes possible an estimate of the effect of sedimentation in capacity on the reservoir and identifying other sources of erosion such as soil erosion from neighboring areas. There are many advantages of this approach. With a DoT, one can cover vast areas very quickly and efficiently in order to collect data from quite some points in the reservoir and provide much more information on sedimentation patterns compared to the older methods. The AI algorithms will also better analyse such data compared to manual methods and facilitate further accurate assessments. Regular monitoring identifies sedimentation problems early on; and timely intervention measures save the reservoir's capacity and hydropower generation from significant losses.

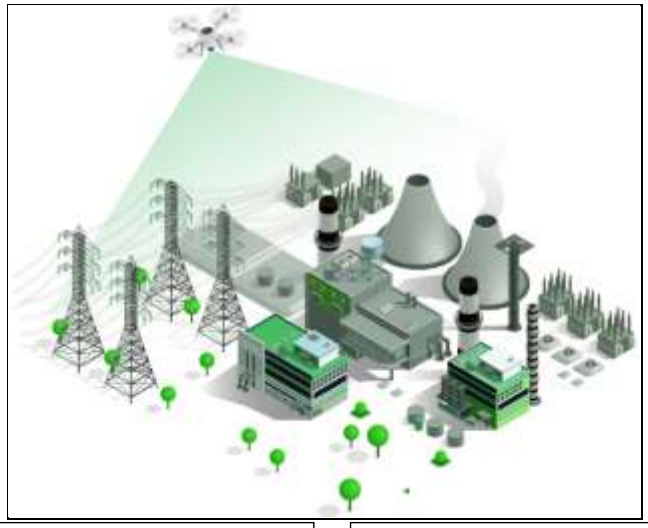
The automated methodology cuts down the necessity to have costly and time-consuming surveys done manually. Hence, it has important cost-saving benefits. Moreover, the technology will help identify sources of erosion as well as sediment pollution through observing patterns in sedimentation. This is essential for environmental protection and the management of water resources. An AI-powered DoT is to be used for sedimentation monitoring of reservoirs, ensuring an appropriate water storage capacity, managing water quality and sustainable resource management of water. One can use it to monitor the effects of sedimentation on hydropower generation, thereby allowing the operators to plan maintenance or dredging operations that would optimize energy production. It can further be used in monitoring environmental changes, source identification of sediment pollution, and assessment of the impacts of sedimentation on aquatic ecosystems. It can even be used in flood control, monitoring sedimentation in rivers and floodplains to assess flood risk and work out mitigation measures.

II.1.4 Thermal Power Plants

While renewable energy sources are gaining prominence, thermal power plants continue to play a significant role in electricity generation. AI-powered drones are proving to be valuable tools for inspecting and maintaining these facilities, ensuring their safe and efficient operation. Fig.7 shows the AI Powered DoT for thermal power plant inspection system.

II.1.4.1 Chimney Stack Inspection

This drone flies hundreds of feet in the sky. In the foreground stands a gigantic industrial chimney stack. Such inspection is reality nowadays with the advent of AI-powered Drone of Things or simply DoT for short. These drones, mounting high resolution cameras and thermal imaging sensors, revolutionize how to inspect and maintain these critical structures venting exhaust gases and pollutants emanating from power plants and industrial facilities. A comparison between the features and systems above with classical approaches such as scaffolding or rope access techniques yields that cost-effectiveness is combined with minimum safety risk for the inspectors. AI-powered drones end the risks of flying this way by offering a safer and more efficient alternative. It can quite easily access



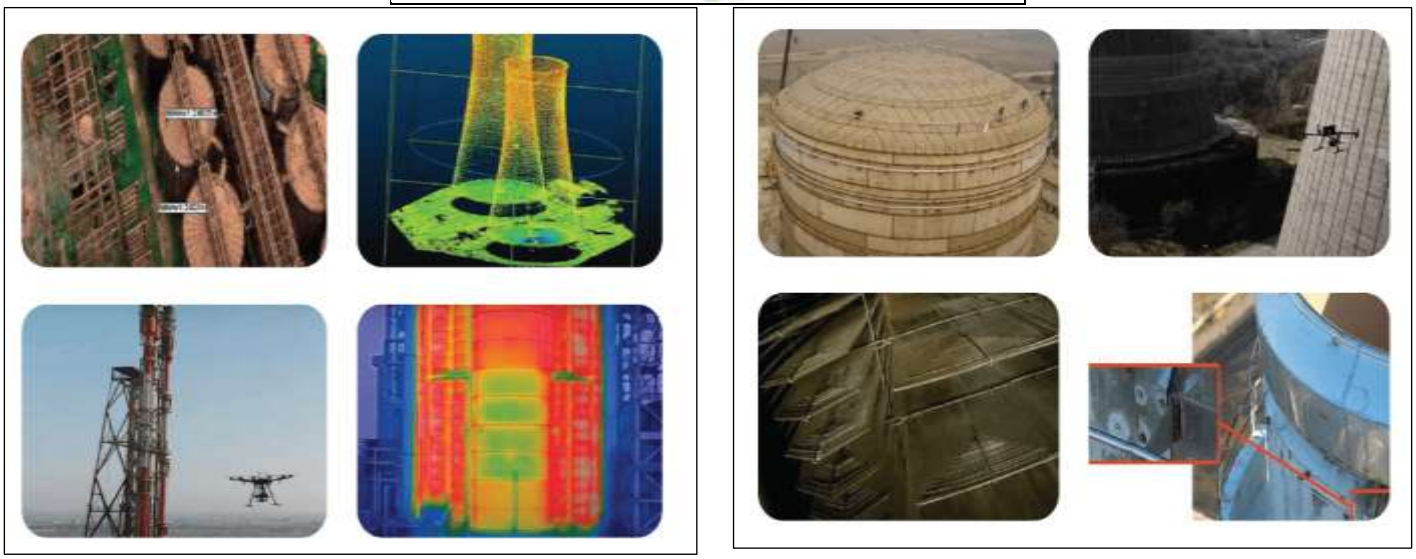


Figure 7: AI Powered DoT for thermal power plant inspection system.
Source: [19].

Inaccessible areas and capture highly detailed images of the chimney's interior and exterior surfaces, encompassing the condition of the brickwork, mortar joints and attached platforms or ladders. The drones' thermal cameras pick on temperature differences that allow it to detect problems such as the possible failure of insulations, structural damage, as well as areas where heat is leaking out and which generally hinder efficiency. It is the advanced computer vision algorithms in analysing the visual and thermal images that pose the power with this technology. The AI algorithms will automatically identify all possible defects including crack, spalling (where concrete flakes off), corrosion, as well as blockages. They can even estimate the degree of these defects, which is priceless information when it comes to prioritizing repair work. The AI system then produces comprehensive reports with detailed visual documentation, thermal maps and precise location information for any defects found. This means engineers and maintenance teams can rapidly assess the chimney stack's condition and plan necessary repairs in advance[19].

This technology has a number of benefits. There is improved safety: because inspectors would not have to climb tall constructions or work at heights to inspect. Efficiency is improved by making inspections automated, hence much faster than other methods. Lastly, accuracy is improved in that AI algorithms can detect minor defects that, through human inspection, would be missed. Further, it reduces the cost of inspection as it does not require scaffolding, rope access or specialized equipment. Lastly, it provides detailed documentation that is very helpful for maintenance planning and regulatory compliance. AI-based DoT chimney stack inspection is widely employed in several sectors of business. This process may be utilized for inspecting the chimney stack in thermal power plants to ensure efficient and safe working. Stacks mounted in industrial plants, refineries etc can be scanned to study the emission of smoke and dust as part of environmental pollution. Ancient buildings with old chimneys and towers also can find support in this technology to ensure their safety and durability.

II. 1.4.2 Cooling Tower Inspection

Regular inspection is required for cooling towers, which are considered the means of dissipating excess heat in power plants and other industrial processes. Traditionally, it involved manual inspection, an agonisingly slow, dangerous and limited process. Drones with AI are fresh in this game and remain safer, faster, and comprehensive when inspecting a cooling tower than any other inspection method in existence. With a high-resolution camera that has a thermal imaging sensor, this drone manages to see every nook and cranny inside the cooling tower as it creates very specific images of the shell, the fill material, the drift eliminators, the piping and even nozzles. It ends up creating a total visual analysis of the structure and its parts that might exist internally and externally. Thermal cameras can detect temperature changes, thereby exposing issues such as blockages, leaks or inefficient cooling, along with even damage not visible to the naked eye. In general, there is true magic in high-tech vision algorithms analysing detailed visual and thermal images.

This AI can itself automatically identify such defects as cracks, corrosion, damage to the fill material, and even biological growth. It also rates the level of severity of these defects, thereby providing essential prioritization for various types of maintenance. The AI will generate reports containing high detail visual documentation and thermal maps, as well as the correct location of any issues that may be detected. This allows engineers and the teams responsible for the maintenance to quickly determine the condition of a cooling tower and what is needed as far as repairs or cleaning. In comparison with other methods, this has a lot of advantages: it is safer because inspectors do not have to climb tall structures or go into confined spaces; it is efficient by being automated, meaning the inspection might be much faster compared to traditional methods; and it is more accurate by having AI algorithms look for defects that are nearly not noticed during inspection on the human body. It also reduces inspection costs since the requirement for scaffolding, rope access, or specialized gear is eliminated and by locating and remedying defects impacting cooling effectiveness, it can help optimize the performance of the cooling tower along with total operational efficiency.

II.1.4.3 Coal Stockpile Volume Estimation

Estimation of the amount of coal in stockpiles is still essential for power plants to manage fuel supplies and work effectively. This used to be done by traditional measurement and estimation through aerial photos, which is a quite inaccurate procedure and time-consuming as well. The use of drones in measuring coal stockpile volume is changing with AI. These drones are equipped with LiDAR sensors that work based on the principle of how laser pulses are emitted and measured by the time travelled after its reflection from the coal pile. It generates a very dense 3D point cloud and, in essence, gives a digital representation of the stockpile's surface. To eliminate errors, the GPS and inertial measurement unit of the drone puts this point cloud precisely within a real-world coordinate system. Noise and outliers are usually in the raw LiDAR data and this is filtered by AI algorithms for clean and accurate representation of the surface.

Then, the AI classified the points in the point cloud into either ground or stockpile to separate the coal pile from the surrounding terrain. AI develops a 3D surface model of the stockpile of coal as portrayed in filtered and classified data. A volumetric algorithm used by specialized software computes the volume of coal stockpile on the basis of such a 3D model. Multiplication of the volume with the bulk density of coal, found from the lab test or history, will yield an estimate for the actual mass of coal. The 3D models developed can be visualized and analysed with the help of specialized software to gain insights into the shape and volume distribution of the stockpile. The AI system automatically generates reports containing all volume calculations plus other relevant data that can be linked to any given inventory management system, and such data facilitates timely tracking and analysis [19].

II.2 TRANSMISSION SYSTEM

II.2.1 Autonomous Inspection and Monitoring of Transmission Systems

Power transmission systems inspection and maintenance are becoming new faces of AI-driven drones; their activities help run processes more efficiently, safely and reliably. Here's a closer look at how intelligent aerial robots is revolutionizing the monitoring of transmission lines, towers, and substations [6].

II.2.1.1 Transmission Line Inspection

The inspection of power transmission lines, those critical pathways delivering electricity from power plants to our homes and businesses, is revolutionized with AI-powered drones. Traditional inspection methods are slow, expensive and very limited in access to remote areas and difficult terrain. Drones with advanced sensors and AI algorithms improve efficiency and enhance safety while providing a much more complete solution. They will take cameras with higher resolutions that will provide high-resolution images of the transmission lines, including the conductors-that is, the wires-themselves, insulators that support the wires and supporting structures. All these will give a visual record of the line's condition, thus making it possible to identify the possible problems. They also have thermal cameras that can spot variations in temperature along the lines. These can highlight hot spots which are an indication of loose connections, corrosion or other issues that could lead to a loss of power or even fires.

Real 'power' lies in those AI algorithms that look at the visual images and the thermal images together. Such algorithms can automatically detect many types of anomalies, including broken conductors, damaged insulators, vegetation creeping onto lines, and loose or missing hardware. This automation does not involve the manual inspection of every single image but instead saves time and is more accurate. Subsequently, with a reasoned conclusion, the AI system generates comprehensive reports along with visual images documenting all the anomalies detected, including thermal maps and the precise location of those anomalies. The maintenance teams would quickly be able then to estimate the condition of the transmission lines and work on a priority list of problems, first those most crucial. This technology has many positive factors. It speaks of safety as it does not require labourers to climb high towers or have access to places that may be hazardous.

It ensures efficiency as inspection is automated, which saves much time from inspection procedures compared with the conventional methods. It also promotes precision because AI algorithms can identify defects that may have been missed by human inspectors. Lastly, it reduces downtime since proactive maintenance becomes effective and prevents costly interruptions in electricity supply. The last is that inspection costs are avoided, and the helicopters, ground patrols or any other special gear can be excluded. AI-powered drone inspections have a number of applications, such as routine inspection, emergency response after the storm or some other event, vegetation management, and construction monitoring. This technology is quite fundamental to the reliable delivery and efficient transmission of electricity with increased sustainability and resilience in the energy grid. Figure 8. Shows the DoT inspecting a power line(a), a close-up image of a healthy insulator(b), and a close-up image of a damaged insulator with cracks(c) respectively [20].

II.2.1.2 Tower and Substation Inspection

Inspection of transmission towers and critical substations controlling electricity always requires high vigilance for ascertaining the reliability and safety of our power grid. Traditionally, inspections were done through manual climbing of towers as well as close-up observation of substation equipment. Though this practice was time-consuming and not cost-free, it also posed a hazard to workers. However, AI-powered drones transform that very landscape with a safer, more efficient, and more all-inclusive approach towards checking the towers and substations. With LiDAR sensors, these drones function as flying 3D scanners that create highly accurate 3D models of towers and substations. It helps inspectors to scan the very structures virtually from all sides and capture their very precise dimensions and geometry in detail. High-resolution cameras mounted on drones take extremely detailed images of each and every aspect of the substation, from the insulators and steel structures to the conductors and other equipment at the substation. Another layer of information which can be added to the thermal cameras installed on the drones is temperature variation detection in equipment and connections, indicating possible overheating or even not tightened connections.

The real power though is realized with advanced computer vision algorithms which analyse not only visual and thermal data but also LiDAR data. AI algorithms work like expert inspectors to automatically identify nearly any defect or anomaly. They can detect structural issues such as corrosion, cracks, and loose bolts in towers and on substation equipment. They may also report equipment irregularities such as overheated transformers, cracked insulators, and malfunctioning circuit breakers. In addition, they can detect development encroaching too close to towers or substation equipment, potentially damaging them or interfering with their operation. The AI system then produces comprehensive reports, which also include detailed visual documentation, thermal maps, 3D models and precision location information for any item found. This allows engineers and maintenance teams to quickly assess the towers and sub-stations so that restoration work is prioritized to address the most critical issues right away [21]. This technology offers many benefits. Firstly, safety is increased because workers are not asked to climb tall towers or to approach high-voltage equipment.

It increases efficiency due to automated inspection processes that would definitely be much faster than by people alone. Accuracy is improved in these inspections as the AI algorithms can detect minor defects that might be unnoticed by human inspectors. This also saves on inspection costs because of no manual climbing, special equipment and no need for extended time schedules; by promoting proactive maintenance, it prevents costly repairs and outages from popping up without warning. With these visual, thermal, and LiDAR data sets, more comprehensive insights about the tower and substation conditions are brought into the decision-making process for maintenance. AI-powered drones inspect towers and substations for routine monitoring, assessment after storm damage, security surveillance and construction monitoring. This technology is indispensable in the quest to ensure reliability and resilience of the power grid while bringing the energy future one step closer to being cleaner and more reliable.



a)



b)



c)

Figure 8: (a) a drone inspecting a power line (b) a close-up image of a healthy insulator (c) a close-up image of a damaged insulator with cracks.
Source: Authors (2026).

II.2.2 Distribution System

AI-powered drones are revolutionizing how we inspect and maintain the vast network of power distribution lines that bring electricity to our homes and businesses. These drones, equipped with advanced sensors and AI algorithms, offer a safer, more efficient and more comprehensive solution compared to traditional methods like ground patrols or helicopter inspections[22].

II.2.2.1 Power Line Inspection

AI-powered drones are changing the process of inspecting power lines. The lines that provide the electricity to our homes and businesses are inspected using advanced cameras and thermal imaging sensors by these drones. Flying electricians, these drones carry out meticulous examinations of each line in distribution. Detailed images taken include the wires, the insulators and even the supporting structures for all lines. The thermal cameras will measure temperature variations and highlight probable problem spots like loose contacts or corrosion. But what is the real magic - advanced computer vision algorithms are used to analyse those pictures. The AI can recognize a very broad spectrum of anomalies - damaged equipment, vegetation overgrowth, violation of rights of way, including the exact pinpointing of where the power outages happened so that crews can regain power to those areas efficiently.

This auto-analysis saves time and ensures accuracy as compared to human inspections. The AI then produces reports that are fully detailed with visual documentation, thermal maps and location information for any issues that are identified. This makes it easier for maintenance teams to quickly evaluate the situation and prioritize repairs so that critical problems are addressed first [21]. This technology provides the following: safety that is eliminated since workers may not require climbing poles, nor enter hazardous areas during inspections, reduces cost to be many times faster for inspections. Improves in precision given the AI could find and pick out better defects often missed by even the expert human inspectors and reduces downtimes because it uses proactive measures for maintenance - no probable chance of discontinuity within the electricity supplies. This further saves the cost of inspections due to the removal of ground patrols, specialized equipment or extended downtimes.

II.2.2.2 Grid Mapping and Planning

AI-powered drones level the playing field, as it were, when it comes to mapping and planning distribution networks. High-resolution cameras, LiDAR sensors and thermal imaging cameras mounted on drones might be said to constitute 'flying surveyors' gathering a treasure trove of data regarding the grid as well as its surroundings. They create very sophisticated models of the 3D network, identify poles and transformers and can even detect problems such as damaged equipment or vegetation encroachment. Then, it goes through advanced AI algorithms, creating a full digital twin of a distribution network. Digital twin provides an in-depth evaluation of the grid mapping of the connection between elements, the analysis of sag and clearance over the power line and even identification of specific locations with a risk from vegetation. But not just in mapping- planning and optimizing the grid is much easier through AI-powered drones.

They can analyze load patterns and predict future demand. Based on this, they can highlight the areas that require expansion or up-gradation within the grid. They can also decide the location of new transformers and substations by working on terrain, population density, and load patterns and also with that new focus towards renewable energy, these drones would be able to scan the possibility of incorporating photovoltaics and wind energy into the grid, as well as the best locations for their incorporation. It has some advantages. It enhances accuracy, efficiency and safety while costs are reduced in comparison to traditional methods. However, the most important thing is that it gives data-driven decision-making opportunities in planning and optimizing grids with valuable insight. This increases the resilience of the grids that can better adapt to future challenges: increasing demand and renewable energy sources [15].

II.2.2.3 Last-Mile Delivery

AI-powered drones change not only the way we inspect and maintain power infrastructure but also the way we respond to emergencies and ensure speedy repairs, particularly in remote and challenging locations. And so, when talk is about "last-mile delivery," AI-powered drones are the game-changers. Imagine a critical piece of equipment fails in a remote power substation high up in the mountains. Using traditional methods, getting a spare there might take hours or even days, leaving people without power for extended periods and causing immense economic loss. That's where AI-powered drones come in. They can be sent over to the location fast, where that essential spare part could be dropped directly into the substation, finding easy routes over treacherous terrains and thus reducing downtime. This reduces disturbances to customers and saves on costs incurred in repairing losses and unproductive time. Besides, when nature does its thing, like destructive weather conditions rendering roads impassable, drones become a very important asset. They can help give a bird's eye view of the disaster situations, and this gives the fastest possible way of establishing the damage areas and which areas require urgent repair. They can carry vital equipment also and hand-over needed machinery to restoration teams in inaccessible areas so they can restore electricity and all crucial utilities much quicker than would otherwise be the case.

III. AI APPROACHES FOR DOT APPLICATIONS

III.1 COMPUTER VISION

Computer vision is thus providing drones with smart eyes that can let them see the world, look at things, and understand what they are seeing. In essence, it involves data extraction with meaningful information taken from images and videos captured by cameras mounted on drones. This creates opportunities to apply drones to a variety of power infrastructure management applications. In many cases, the images from the drone are not suitable for analysis without some fine-tuning. That's where image processing comes into play. It's basically optimizing a photo before uploading it to social media. Such techniques as noise reduction, contrast adjustment, and sharpening enhance the quality of images so that the computer may identify and analyze the relevant features[23-26]. Once prepared, computer vision algorithms are further used to identify parts of specific interest. For power lines, that would mean identification of insulators, transformers, vegetation, and even the power lines themselves.

Algorithms such as YOLO, Faster R-CNN, and SSD are well-suited at that because they can quickly and precisely locate those parts in that image. Beyond object detection, computer vision also helps in separating images into meaningful regions. For example, it is an outline of different parts of a picture. Example: It can isolate the damaged portion by using Mask R-CNN and U-Net algorithms if the image of a broken insulator is captured by a drone. This enables one to establish the amount of damage and therefore order repair. Another massive component of computer vision would be image classification. In this case, the classing of an image into a certain pre-defined class is done. For instance, an algorithm may classify a picture of power line conditions as "normal", "minor damage" or "severe damage." This enables one to understand the situation much faster on the site. Computer vision allows features to be extracted from images, such as edges, corners, and textures. These can further allow the identification of patterns or anomalies that may imply a trouble.

For example, changes in the texture of one of the elements of a power line could indicate corrosion. Computer vision can thus generate 3D models for the components of the power infrastructure by combining various images captured from different angles. The detailed and wide scope view offered by such a model helps in an effective inspection and analysis process. For the case of power line inspection, computer vision allows a drone to automatically detect most kinds of defects, ranging from broken insulators, to damaged conductors, encroaching vegetation, to rust. This really accelerates the inspection process and minimizes the possibility of human error. Besides, computer vision is used in monitoring substations; identifying possible dangers; damages created after storms; even tracking vegetation near power transmission lines. The computer vision is one critical technology that empowers the drones to effectively monitor and manage the power infrastructure. It allows drones to see and perceive the world around them by providing great insights to increase power systems efficiency, reliability, and safety.

III.2 MACHINE LEARNING

Machine learning is giving a brain to drones, which enables it to learn and adapt as it goes. Drones need not be programmed for every situation because they themselves can analyze data, determine patterns and make decisions when equipped with machine learning. This learning capability is revolutionizing how we maintain and manage power infrastructure[27-29].

In machine learning, the most common technique used is supervised learning where an algorithm is trained on a set of labeled data. It may be considered as giving the drone lots of examples of what good equipment is and what bad is, and subsequently letting it learn to discern the difference. Through this, the drone will be able to decide outcomes, such as how much useful life remains in a transformer or what condition the power line is in. It is further divided into regression, which predicts a continuous value like temperature, and classification, which predicts a categorical value like "good", "fair" or "poor."

Unsupervised learning deals with unlabeled data. Here the algorithm is asked to find hidden patterns and structures in the data. For instance, it can analyze the pattern of drone flight patterns and determine anomalies such as when a drone flies from its usual route. This might appear as a problem that requires further investigations. Techniques for unsupervised learning include: Clustering where similar data points are often grouped together and dimensionality reduction simplifies the data reducing the number of variables. Reinforcement learning is really teaching a dog new tricks through rewards and punishments. It learns based on interaction with the environment and receives feedback for its actions. In drone usage, it could be related to optimizing routes during inspection: the algorithm will test different routes, which it learns are most efficient and effective based on the feedback it receives.

The scope of applications of machine learning in power infrastructure management is broad. For example, a drone scanning sensor data to predict equipment failure before its event occurrence allows for proactive maintenance without having costly outages and ensuring reliability in the grid. Another application of machine learning could be anomaly detection, where it detects the deviation in patterns within sensor data or images, which are usually unusual and might signal a problem. This implies that there are irregular changes in temperature inside a transformer, abnormal vibrations on a power line or even shady activities that might indicate energy theft. Further, machine learning can optimize the operation of drones. For example, it can decide on optimal flight paths, schedule inspections and also rationalize resource use. This way, drones are optimized in order to maximize potential impact on power infrastructure management.

III.3 DEEP LEARNING

Deep learning is kind of advanced artificial intelligence inspired by the human brain. It is developed through artificial neural networks - very complex structures, modelled closely to how our brains operate when processing information. The networks have many layers in them, and hence, they can learn deep and complex patterns and representations from gigantic sets of data. This makes deep learning particularly suited for processing unstructured data such as images and videos, which are also abundant in the drone-based inspection of power infrastructure [30], [31]. Among the primary architectures for deep learning are Convolutional Neural Networks (CNNs). CNNs are designed precisely to process grid-like data, such as images. They are highly competent in object detection, image classification and segmentation tasks. This means they are applicable for visual data analysis captured by drones during power line inspections. For example, a CNN can be trained for recognizing different power line components: insulators, conductors, and towers, then detecting possible anomalies or defects that may exist.

Another basic architecture is Recurrent Neural Networks (RNNs). RNNs are best suited to sequential data, like time-series data from drone sensors. This really makes them valuable for predictive maintenance and anomaly detection. For instance, an RNN could analyse vibration data collected over time from a transformer and learn to predict potential failures based on slight changes in the vibration patterns. GANs represent a quite recent development within deep learning. They include two collaborating neural networks: the first builds new data while the other attempts to say whether its generated or real. Based on this innovative architecture, GANs can efficiently produce synthetic data that might closely resemble true data, and in terms of DoT, can be adopted to generate datasets for other AI models, especially for scenarios of scarce or pricey real world data. Some of the following benefits of deep learning can be leveraged for drone-based power infrastructure management: it enables much more accurate and robust image analysis especially in difficult conditions such as low illumination, or complicated backgrounds.

This is important for dependable classification and detection of defects. Deep learning also automatically performs feature extraction; and the task requires no time-consuming effort of manual feature engineering, again requiring significant domain expertise. Deep learning also stands out in data fusion-the process of integrating data coming from different sources to get an even better view of the power infrastructure. For example, it can integrate camera visual data, infrared thermal data and LiDAR data to create an even more detailed 3D model of a power line and identify risks more accurately. In essence, deep learning equips drones with a higher level of intelligence than simple algorithms, allowing them to analyse complex data, learn from experience, and make decisions. This increases the efficiency, accuracy and reliability of power line inspections, predictive maintenance and other critical tasks in power infrastructure management. With the power of deep learning, we are moving toward a more resilient and sustainable power grid.

III.4 NATURAL LANGUAGE PROCESSING

NLP is that subset of AI focused on enabling computers to understand and interpret the human language and eventually generate it. It is essentially teaching a computer how to read, write and talk-to bridge the gap as they communicate with each other. The DoT relies on NLP for automated, task-related use involving human languages-for the generation of reports and communication[32-34].

Another impressive technique applied in NLP is text summarization- the reduction of very extensive data sets to more concise summaries. This is highly applicable in DoT where tremendous amounts of data are collected by the drones during the inspection. Human operators gain better understanding and take appropriate actions on the information provided by automatically generated key findings and recommendations summarizing inspection reports from NLP algorithms. Sentiment analysis is another useful NLP technique. This would determine the emotional tone of a text and thus may be useful to find the public sentiment over the proposed route of the power line or to analyse customer feedback.

Machine translation translates reports and other documents from one language to another for further effective communication and collaboration across various regions and countries. Speech recognition endows the ability to perceive and understand voice commands in drones for more intuitive yet effective control. It can be especially useful, especially when manual control cannot be achieved or even is impossible. Natural language generation helps the drones to produce human-readable reports from the data it has collected. This automatically made report writing since the effort and time of human operators will be spared in the process. DoT uses NLP mainly for report generation purposes. It can automatically create a report of a very detailed inspection findings and recommendations as well as visual representations of data. Of course, it saves time and eliminates human error or variation in reporting. NLP allows drones to make use of the natural language to communicate with the human operators, thus enabling more effective cooperation and troubleshooting. Using NLP, apart from answering the given questions, it can be used in the retrieval of information where drones can fetch the respective meaningful information from various documents and databases to further aid the decisions to be made, problems solved etc.

There are, however, a number of considerations to be taken into consideration when applying NLP in DoT; importantly data quality, which is critical since NLP models rely mostly on training based on the available data assuming that such data would be right and representative. Model explainability is also crucial for building trust and making the decisions the AI system makes transparent and understandable. Ethical considerations, such as addressing potential biases in algorithms, are also paramount to ensure that AI is used responsibly in the management of power infrastructure. Finally, resources needed for computational machines must be brought in because training and deploying NLP models, especially deep learning models, can require significant computing power. With effective use of NLP and other AI approaches, DoT can significantly enhance power infrastructure monitoring, maintenance, and management to improve efficiency, reliability and safety toward making the power grid more resilient and sustainable.

IV. IMPLEMENTATION CHALLENGES AND CONSIDERATIONS

Implementing AI-powered Drone of Things (DoT) systems in the power sector comes with its own set of challenges and considerations[35-38]. Here's a breakdown:

IV.1 REGULATORY FRAMEWORKS AND AIRSPACE RESTRICTIONS

It is an excellent benefit when it comes to using drones in the case of power line inspection, but with unmanned aircraft comes a great deal of regulatory considerations. Just as manned aircraft fall under a very complex web of rules designed to ensure safety and security, drones, too, require such regulation in place. Regulations vary from country to country but are far more strict for operations beyond the line of sight of the visible work, where the aircraft flies out of the operator's line of sight. This is the kind of operation often called upon in the inspection of long stretches of power lines in less expensive and efficient ways. The activity is very time-consuming, as most of it involves careful planning and attention to definite requirements for obtaining all permits and licenses for BVLOS operations. Another issue is airspace restriction. The power infrastructure sprawls over vast geographic distances, for example areas adjacent to airports, military bases or any other sensitive point at which the flights of drones may be either restricted or generally prohibited. In this regard, their flight paths may have to be changed or some alternative means applied for inspection purposes.

One needs to properly plan crossing these restrictions and coordinate duly with authorities while, in some cases, obtaining special permissions. Besides such deep concerns about safety and security, even very elementary privacy issues have to be factored in. For instance, even normal applications like line inspections with drones would provoke privacy issues over data collection over some private properties. Responsible acquisition of data will require guidelines and clear-cut protocols that ensure proper consent, anonymization of sensitive data and transparency over uses of collected data. It would be basically important to find a balance between the benefits of drone inspection and their corresponding infringement on privacy rights so that the said technology would also be done ethically and accepted by the public. In summary, though, drones may become a very potent tool for inspection of power infrastructure, but their operation is to be done in tandem with a framework of regulations and ethics. Understanding of drone regulations, airspace restrictions and privacy considerations becomes very important in safe and responsible and effective power sector deployment.

IV.2 INTEGRATION WITH EXISTING POWER INFRASTRUCTURE AND DATA SYSTEMS

Integration of advanced AI-powered DoT systems in the existing power grid infrastructure is beyond the simple plugging in of new technology into existing hardware and software networks of managing the power grid. Another major impediment the new DoT systems may face is compatibility with the existing infrastructure with regard to data formats and communication protocols. This could turn out to be an enormous job on translation of data, establishment of secure communications, and initiation of data flows between these disparate systems. Without compatibility, precious data from drones may not flow effectively into the power grid management systems. Another important challenge is that several power companies are holding onto older legacy systems, which, although working correctly, may not be suitable for the quantity of information being churned out by modern AI and drone technologies. They will probably not have sufficient processing speed, storage capacity, or flexibility to integrate with those new systems.

Upgrading and adapting these legacy systems, although possible, can both be expensive and time-consuming and require a lot of investment and potentially disrupt working patterns. In addition, DoT systems produce a large amount of data. Inspection by drones with high-resolution cameras and various sensors would generate significant amounts of data. Such data needs a robust infrastructure for storage, processing and analysis. This would demand enough storage capacity but also includes efficient data processing pipelines and analytics platforms that could extract meaningful insights from otherwise raw data. Without proper data management, the valuable information collected by drones might get lost in the flood of data, thus hindering the potential benefits of the DoT system. Integration of AI-DoT systems into the existing power infrastructure will require careful planning and consideration regarding compatibility issues, legacy systems, and data management challenges. Building solutions that overcome these challenges can be true aids in unlocking the real potential of DoT technology and its successful implementation in the power sector.

IV.3 ACCURACY AND RELIABILITY OF AI ALGORITHMS

The quality of the algorithm used will determine how effective AI-powered drone inspections would be. Algorithms may be strong but not a hundred percent accurate always and in many cases, performance is also influenced by lots of variables. The quality of training data is another such major variable. Because the AI algorithm learns from the input data, if that input data happens to be wrong or incomplete, the outputs made and decisions taken by the algorithm will be untrustworthy. For instance, taking the training of an algorithm with images of pristine power lines, the same algorithm, while facing tiny defects or anomalies on real world power lines, fails to detect them. Thereby, it is essential that the training data prepared should be the right way with diversity, representativeness, and accurate labelling for reliable AI models.

Environmental factors also greatly influence the precision of the AI algorithm, especially those using computer vision. Weather conditions, lighting and image quality can all affect how well an algorithm could interpret visual data. For instance, an algorithm trained on clear, sunny images may find difficulty identifying defects in images captured under foggy or overcast conditions. Other problems include insufficient illumination or casting shadows. This means missing information and a rating which might be incorrect, so correcting this issue demands developing algorithms insensitive to variations in environmental conditions that could get relevant information from poor-quality images. The algorithm itself should be resilient. In addition to quality and the quality of its data, an AI algorithm needs to be reliable under unexpected events or a change in conditions.

It needs to have the ability to handle noisy data, learn about new situations and gracefully recover from exceptions. Testing and validation must be exhaustive to establish consistency and accuracy over a variety of scenarios. This will involve testing the algorithms against different datasets, simulating various environmental conditions, and comparing their performance with established benchmarks. It is only when one gets the accuracy and dependability of AI algorithms will they be able to perform any operation using drones for checking the power lines. They should focus on the proper quality of training data that does not have environmental factors working as obstacles and test and verify them properly to ensure it's robust and reliable. Considering these requirements, we will be able to design trustworthy AI systems with accuracy, enhancing the safety and efficiency in the management of the power infrastructure.

IV.4 DATA SECURITY AND PRIVACY CONCERNS

With the advent of AI-powered drones in power infrastructure inspections comes a critical responsibility: the protection of the sensitive data that these drones collect. That information not only is very detailed about the power grid itself but may also accidentally capture images of private property and individuals. Protecting such information from unauthorized access, misuse and breaches is important. Data breaches can be very crucial because they might expose critical infrastructure to sabotage or disrupt operations, which affects sometimes national security. Cyberattacks launched against DoT systems can steal sensitive information, tamper with inspection results, or even hijack a drone control system. To this effect, proper cybersecurity will involve efficient encryption, access controls and intrusion detection systems and prevent unauthorized access into the system. Regular security audits and vulnerability assessments are also provided to identify potential weaknesses in the system. This is aside from security and there's the critical aspect of data privacy. Clear policies and procedures should make sure that the data which is collected is used correctly and stored without violating individuals' rights as well as privacy regulations.

It relates to obtaining any kind of consent for collecting data, anonymising sensitive information or restricting a certain amount of access from people who are only authorized. Observing data protection principles enhances public confidence and ensures the ethical use of DoT technology. Lastly, determination of clear data ownership is necessary. Determination of the ownership of the information obtained by DoT system-the power company, a technology provider or third parties-would have legal operational implications. Clear ownership in place will ensure that the data managed are kept secure and under control of those best placed to comply with applicable requirements. Determination of the right ownership also will resolve every possible argument over the data at hand. In conclusion, data security and privacy concerns have to be met in a responsible and effective manner before AI-powered drones can take to the skies in the power sector. Data protection, strict security measures, and respect for ethical ways of handling data will give us the power to use these technologies in our favour, while keeping sensitive information secure and under public confidence.

IV.5 LIMITATIONS OF AI-DOT SYSTEMS IN HARSH ENVIRONMENTS

The AI-based solution is extremely promising, though environmental interference is the reason that leads to its decreased efficiency, so DoT systems in particular are of a real necessity to further continue the scope of research and development in areas where it improves the level of robustness and adaptability.

IV.5.1 Weather Woes

Strong winds, heavy rain and snow conditions pose a severe restriction to drone activities. Strong winds may make flight unstable and the picture blurred. Rain and snow may reduce the visibility and affect the functionality of onboard sensors. Sometimes extreme weather even makes it dangerous to deploy drones and therefore inspections have to be rescheduled. This may result in increasing latency in identification of potential issues and risk of power outages.

IV.5.2 Electromagnetic Interference (EMI)

Power infrastructure itself can induce EMI which might interfere with the communication system of the drone or its sensor data readings. The high voltage power lines and electrical substations are sources of electromagnetic fields that have impact on the navigation, communication signals and sensor data in the drone. This might result in inappropriate data collection or loss of communication with the operator and even cause safety hazards if it starts interfering with the flight control.

IV.5.3 Terrain Problems

Power lines are hard to navigate and operate over in difficult and remote terrain, such as mountains or forests. The obstructions due to trees, cliffs and undulating terrain may cause obscurations that will make it impossible for the drone to keep a clear line of sight with the lines under inspection, hence making it impossible to carry out optical inspections. The support of Global Positioning System (GPS) navigation may be weak in such environments. Also, poor access to these areas can be challenging to launch, recover, or provide support for maintenance on the drones. Overcoming these limitations would require continued research and development on:

- **Making more robust drones:** Design drones that are more resistant to weather, have stronger communication systems, and advanced sensors that can work well even in challenging environments.
- **Enhancement of Algorithms:** It may focus on enhancing AI algorithms that are less susceptible to environmental interference and which could robustly extract meaningful information from noisy or incomplete data.
- **Improvement of Navigation and Control Systems:** Advanced navigation and control systems which can operate reliably in GPS-denied environments and challenging terrain.

The AI-DoT systems thus can be extended to make them operational even in highly demanding environments through the resolutions of these challenges. The approach towards integration of AI powered DoT systems in power sector calls for multi facets of solution for various associated challenges and considerations discussed earlier. This will require more collaboration, investment, development in research and commitment to best ethical AI practices.

- **The Way of Collaboration:** DoT technology requires collaboration of multiple stakeholders for its effective deployment. Power companies, technology providers and the regulatory bodies need to join hands to develop industry standards, share best practices and address problems related to regulatory hurdles in the implementation of DoT. Such collaboration will ensure safe, responsible and beneficial application of DoT systems toward both the power industry and the public.
- **Infrastructure Investment:** Huge investments in infrastructure would be necessary to support DoT systems. This includes the need for strong data management systems that can handle the vast data generated by drones, development of strong cybersecurity measures to secure sensitive information, and developing AI algorithms that can provide accurate analysis and reliable predictions. Investment in training and development for personnel is also needed so that the workforce may deliver the operating and maintaining skills associated with these systems.

Continuous research and development needs to be applied in AI so that this area remains highly evolved while developing its systems with strong robustness, reliability and accuracy of the AI-DoT. That means designing better and advanced AI algorithms as well as advancing the capabilities of hardware components and sensor types on the drones. Environmental adversities pose another threat that is bound to happen with any developed system or structure; the only antidote for continuous innovation so DoT stays abreast.
Ethical considerations come to the forefront, so ensuring the responsible use of AI builds public trust and makes way for the acceptance of DoT technology. There must be consideration of issues surrounding data privacy, algorithmic bias and even job loss. Thus, clear ethical guidelines have to be developed; the public has to be assured by means of openness; public discourse has to be made an integral part of confidence-building in this technology. By embracing the multi-faceted approach, the power industry will better overcome all the challenges associated with AI-DoT systems and tap into this transformative technology for full expression. It is bound to lead to safer, more efficient and more reliable power infrastructure for the industries and communities that it shall service.

V. FUTURE TRENDS AND OPPORTUNITIES

The Drone of Things is rapidly evolving, driven by advancements in AI, communication technologies and sensor capabilities. Here are some key future trends and opportunities that will shape the future of AI-DoT in power infrastructure [39-42]:

V.1 IMPACT OF 5G AND EDGE COMPUTING ON DOT PERFORMANCE

This integration of 5G networks and edge computing could be taken as the future direction to revolutionize DoT system capabilities in terms of real-time data processing and autonomous drone operations. Envision a drone flying through networks of power lines inspecting them and taking in images as well as sensor data from the flight. All captured data in 5G can be transferred nearly in real time to a central control center where AI algorithms could analyze it for defects or anomalies. It's just about real-time transfer; the bandwidth is high and the latency is low. An immediate analysis can occur due to faster response times. This would alert the operators, who would immediately take necessary actions and thus prevent an outage. Such a response capability would be very crucial in an emergency situation because the damage caused could be minimized with prompt assessment and action and quick restoration of power supply. In this scenario, with edge computing, the power processing has brought it one step closer to where the data originated from. No longer need it travel its way through to distant servers located in a cloud for analysis; this can be done within the drone itself or at an edge server located nearby.

That again reduces latency and makes possible near-real-time decision-making. For instance, the drone can automatically change its flight path to take a closer look at a potential defect or even raise an alert if it identifies a critical problem. Increased autonomy in this regard improves efficiency but also allows drones to operate in remote or challenging environments where communication with a central control center might be unreliable. This powerful combination unlocks new possibilities for DoT systems. It allows operators to make immediate decisions to react to potential problems before an outage can occur. It enhances the situational awareness of these systems through the steady stream of real-time data regarding the health of the grid. It enables operators to use resources more effectively by allowing drones to be more autonomous and to make adjustments based on changed conditions. Basically, 5G and edge computing are making drones change from data-gathering robots to the smarter agents of analysis on their own in real time, which marks a tremendous stride toward that future of utilizing drones much more integrally in establishing reliability, resilience and sustainability within power infrastructure.

V.2 DEVELOPMENT OF SWARM INTELLIGENCE FOR COORDINATED DRONE OPERATIONS

Imagine a fleet of drones as taking off not necessarily as separate units but rather as a collective swarm, like a flock of birds or a school of fish. This is the vision about swarm intelligence where numerous drones work in cooperation and interactive communication as if a single coherent entity acting with a common purpose. This methodology could be used to change the way we look at and maintain our critical infrastructure in power line inspection. A swarm of drones allows dividing and conquering where different parts of the power grid will be inspected simultaneously instead of relying on one drone to cross large distances. This saves much time for inspection thus enabling frequent and comprehensive review. Also, the concerted movements of the swarm leave no part untouched hence perfect coverage with reduced chances of missing possible faults. This is particularly helpful in inspecting power lines that traverse difficult terrain or remote areas where access might be limited and coverage hard for a single drone. Cooperative properties of swarm intelligence also open avenues for the execution of more complicated tasks. For instance, in coordinated maintenance, drones may change insulators, particularly damaged ones.

During emergencies, a swarm can quickly assess damage, identify weak points and transmit updates to response teams, facilitating fast and efficient rehabilitation. However, significant advancements in drone technology are necessary to fully realize the potential of swarm intelligence. Communication: Drones must be able to communicate perfectly with each other and share information about their location, findings and tasks. This could require robust and reliable communication networks, perhaps using 5G or mesh networks. Advanced algorithms have to be developed in order for the swarm to coordinate and work together without collision or conflict. The inspiration for this draws from nature, studying coordination of movements of insect or animal swarms, and applies these principles to drone technology. Conclusion Swarm intelligence brings a paradigm shift to the ways drones can execute power line inspection and maintenance in much more effective, total and adaptive ways. In achieving this stable and sustainable energy future, wherein power grid reliability and resilience improvement rest on core-drone swarms cohesively working together, ends.

V.3 ADVANCED SENSING SYSTEMS FOR ENHANCED DATA ACQUISITION

The future of DoT in power line inspection lies in the expansion of the sensory capabilities of these aerial robots. Although high-resolution cameras provide useful visual information, equipping drones with advanced sensors unlocks a new dimension of data acquisition, allowing for more comprehensive and accurate assessments of power infrastructure. For example, hyperspectral cameras capture more information than even a normal camera by penetrating the spectrum of the entire electromagnetic spectrum. That is, the cameras "see" in more detail than the naked eye and can detect little variations in light that identify hidden defects or stress in power line components. For instance, a hyperspectral camera is able to detect signs of corrosion or material degradation invisible to an ordinary camera. In LiDAR sensors, detection and ranging work by transmitting laser pulses into the air and measuring their return, which allows the development of sophisticated 3D maps over the surroundings. This serves the purpose in evaluating whether vegetation has started encroaching into space, pointing out trees whose branches and roots could have reached areas that may jeopardize overhead lines, or in more abstract terms, helps to give clear and accurate models of infrastructure in the form of geometric data to base maintenance on or to test scenarios that would require considerable amounts of resources and times to assess on site.

New, more advanced thermal imaging technologies have been developed to improve the inspection capability. Sensors in these technologies detect the heat signatures, which prove very helpful in identifying overheated components as well as loose connections, signs of potential failure. These are especially helpful in transformers, insulators, and other electric components, where temperature variations may point out the problems. With such advanced sensors integrated with AI algorithms, the whole inspection process of drones reaches another dimension. Data collected from such sensors will be further analyzed by AI to reveal any kind of hidden patterns or anomalies that go unnoticed through a human's eye and insight.In other words, AI-based algorithms will inspect hyperspectral data to indicate minute levels of corrosion; LiDAR data for 3D modeling of vegetation overgrowth, and thermal image interpretation of overheating of components. Advanced sensors' powers combined with AI's analysis ability might well lead to a more informed understanding of the power infrastructure status, which ensures better planning in maintenance, quicker response to potential concerns, and therefore a more efficient, stronger power grid.

V.4 POTENTIAL OF AI-DOT FOR GRID OPTIMIZATION AND RESILIENCE

Undoubtedly, the use of AI drones for both inspection and maintenance is quite beneficial, nevertheless it has other applications as well. This Advanced Intelligent Technology, AI-DoT, is therefore expected to revolutionize the operational capabilities of the grid as well as increase the stability of power systems, resulting in a better energy system than the present. Drones constantly operate using multiple sensors and artificial intelligence to enhance the security level of the power system. In addition, these devices are capable of transmitting data in regards to the power flow, voltage levels and environmental parameters active within the grid in real time, thus creating a more detailed and working picture of the grid’s operation. This data can help understand the range of potential threats, including but not limited to high line loading, transient voltage excursions and dangerous weather patterns. In turn, successful tackling of these issues would allow grid controllers to avert load shedding, improve the operational efficiency of the grid and enhance the management of electricity supply.

The focus of additional enhancements of grid systems with AI-DOT will also touch the aspects of improvement of efficiency of grid management. What is the proper way to achieve this? The situation will be analyzed using statistical and historical evidence. The AI algorithms will be driven to recognize the changes and they will help predict how much the natural loads will increase, which would help the operator in properly allocating the power supply to the system while reducing the outflows of energy. This will also lower the operation cost incurred in the electric system and help reduce adverse effects that are created by the generation of power. In addition to that, AI-DoT will aid in bringing in more resources such as solar and wind energy into the power grid management by keeping track of the resources and adjusting the power grid accordingly. There’s also the backbone of the AI-DoT, which serves the same purpose as any operating system in a cloud-managed service and adds a layer of power network resilience.

VI. CONCLUSION

As detailed in this paper, AI powered drones are set to revolutionize the management and delivery of electric power services. The invention of Drone of Things is in the process of taking place where the demand of power essences has shaped management of power infrastructure using artificial intelligence and drones. Application advantages are many especially in AI-DoT. Tasks such as inspection and maintenance, which would have required manual labor, replace with automation hence improving efficiency. Drones operate in locations that pose a risk to human workers or where it may be too difficult to reach. The knowledge of how to process large amounts of information in a very short time is also an asset of AI. You can be sure that even the tiniest defects will stand out in the images as AI brings modern technology in defect detection. Thus leading to predictive maintenance which focus on performing maintenance activities when the system is likely breaking down averting loss due to service outages hence increasing grid reliability. The AI-DoT has increasing pace of innovation in power sector, further research and development efforts will be required to exploit the full potential of AI-DoT.

This includes development of AI algorithms for complex data processing in changing environments, improvements in drones and many different specific types of sensors used on them, as well as sociological issues associated with data handling and bias in AI solutions. Also, new uses of this technology like grid optimization, emergency response, infrastructure planning, and similar will generate larger revenue for AI-DoT. AI-DoT is more than just an opportunity for improvement in technology; it provides us with an entirely different way of handling and controlling the power system. And in doing so, we are preparing the grounds for a stronger and more integrated power grid in the future. AI equipped drones are not only the tools; they are the active contributors for an advanced, better, cleaner operational energy environment. With the greater demand of electric energy as an essential commodity, it is unavoidable that AI-DoT will be paramount in lessening the pressure on the prevailing infrastructure.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Manjunatha H M and Madan H T
Methodology: Manjunatha H M and Madan H T.
Investigation: Poshitha B and Promodh H K.
Writing – Original Draft: Manjunatha H M and Madan H T
Writing – Review and Editing: Lokesh Mallaiah.
Resources: Swapna H and Aruna B.
Supervision: Manjunatha H M.
Approval of the final text: Madan H T and Lokesh Mallaiah.

VIII. REFERENCE

[1]Matani, Ashok G. "Internet of things and internet of drones in the renewable energy infrastructure towards energy optimization." AI and IOT in Renewable Energy (2021): 15-26.

[2] Pekias, Antonios, George S. Maraslidis, Markos G. Tsipouras, Fotis N. Koumboulis, and George F. Fragulis. "Power Supply Technologies for Drones and Machine Vision Applications: A Comparative Analysis and Future Trends." In Telecom, vol. 4, no. 3, pp. 459-476. MDPI, 2023.

[3] Chehri, Abdellah, Gwanggil Jeon, Issouf Fofana, Ahmed Imran, and Rachid Saadane. "Accelerating power grid monitoring with flying robots and artificial intelligence." IEEE Communications Standards Magazine 5, no. 4 (2021): 48-54.

[4] Shah, Imdad Ali, Noor Zaman Jhanjhi, and Raja Majid Ali Ujjan. "Use of AI applications for the drone industry." In Cybersecurity Issues and Challenges in the Drone Industry, pp. 27-41. IGI Global, 2024.

[5] LNC Prakash, K., Santosh Kumar Ravva, M. V. Rathnamma, and G. Suryanarayana. "AI Applications of Drones." Drone Technology: Future Trends and Practical Applications (2023): 153-182.

[6] Nicolai Iversen, Oscar Bowen Schofield, L. Cousin, N. Ayoub, G. V. Bgel, E. Ebeid. "University of Southern Denmark Design, Integration and Implementation of an Intelligent and Self-recharging Drone System for Autonomous Power line Inspection."

[7] layiwola, Olufemi, Miles Elsden, and Mahmoud Dhimish. "Robotics, Artificial Intelligence, and Drones in Solar Photovoltaic Energy Applications Safe Autonomy Perspective." Safety 10, no. 1 (2024): 32.

[8] https://enterprise.dji.com/news/detail/drones-for-solar-farm-inspections

[9] Henry, Chris, Sahadev Poudel, Sang-Woong Lee, and Heon Jeong. "Automatic detection system of deteriorated PV modules using drone with thermal camera." Applied Sciences 10, no. 11 (2020): 3802.

[10] https://bncfin.com/uav/hyperspectral-and-multispectral drones/#:~:text=Multispectral%20drones%20are%20specialized%20unmanned,images%20across%20multiple%20wavelengths%20of

[11] https://www.edmundoptics.com/knowledge-center/applicationnotes/imaging/hyperspectral-and-multispectralimaging/#:~:text=Figure%201%3A%20Only%20a%20small,visible%20spectrum%2C%20are%20utilized%20in

[12] https://eos.com/blog/multispectral-vs-hyperspectral imaging/#:~:text=Hyperspectral%20imaging%20captures%20ultra%2Dfine,material%20analysis%20and%20target%20detection.

[13] https://en.wikipedia.org/wiki/Hyperspectral_imaging#:~:text=The%20goal%20of%20h perspectral%20imaging,of%20finding%20objects%2C%20identifying%20materials%2C

[14] https://oceanservice.noaa.gov/facts/lidar.html#:~:text=Lidar%2C%20which%20stands20for%20Light,laser%20to%20measure%20ranges%20(variable

[15] https://www.sitemark.com/blog/lidar-a-must-have-for-optimally-designing-your solarsite/#:~:text=The%20drone%20captures%20overlapping%20terrain,to%20create%20a%203D%20model.

[16] https://www.flyability.com/blog/wind-turbine-drone-inspection

[17]Shihavuddin, A. S. M., Xiao Chen, Vladimir Fedorov, Anders Nymark Christensen, Nicolai Andre Brogaard Riis, Kim Branner, Anders Bjorholm Dahl, and Rasmus Reinhold Paulsen. "Wind turbine surface damage detection by deep learning aided drone inspection analysis." Energies 12, no. 4 (2019): 676.

[18] Restas, Agoston. "Water related disaster management supported by drone applications." World Journal of Engineering and Technology 6 (2018): 116-126.

[19] https://garudauav.com/thermal-power

[20] Nicolai Iversen, Oscar Bowen Schofield, L. Cousin, N. Ayoub, G. V. Bgel, E. Ebeid. "University of Southern Denmark Design, Integration and Implementation of an Intelligent and Self-recharging Drone System for Autonomous Power line Inspection."

[21] Luo, Yanhong, Xue Yu, Dongsheng Yang, and Bowen Zhou. "A survey of intelligent transmission line inspection based on unmanned aerial vehicle." Artificial Intelligence Review 56, no. 1 (2023): 173-201.

[22] Hosseini, Mohammad Mehdi, Amarachi Umunnakwe, Masood Parvania, and Tolga Tasdizen. "Intelligent damage classification and estimation in power distribution poles using unmanned aerial vehicles and convolutional neural networks." IEEE Transactions on Smart Grid 11, no. 4 (2020): 3325-3333.

[23] Prebeck, Sebastian, Sathya Ashok, Mounika Vaddeboina, Keerthikumara Devarajegowda, and Wolfgang Ecker. "A Scalable, Configurable and Programmable Vector Dot-Product Unit for Edge AI." In MBMV 2022; 25th Workshop, pp. 1-9. VDE, 2022.

[24] Hassaballah, Mahmoud, and Ali Ismail Awad, eds. Deep learning in computer vision: principles and applications. CRC Press, 2020.

[25] Elyan, Eyad, Pattaramon Vuttipittayamongkol, Pamela Johnston, Kyle Martin, Kyle McPherson, Carlos Francisco Moreno-García, Chrisina Jayne, and Md Mostafa Kamal Sarker. "Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward." Artificial Intelligence Surgery 2, no. 1 (2022): 24-45.

[26] Rajesh, Varadi, and Umesh Parameshwar Naik. "Quantum convolutional neural networks (QCNN) using deep learning for computer vision applications." In 2021 International conference on recent trends on electronics, information, communication & technology (RTEICT), pp. 728-734. IEEE, 2021.

[27] Nichols, James A., Hsien W. Herbert Chan, and Matthew AB Baker. "Machine learning: applications of artificial intelligence to imaging and diagnosis." Biophysical reviews 11 (2019): 111-118.

[28] Momina, Momina, Qurtulen Qurtulen, Hesam Salimi Shahraki, Anees Ahmad, and Zainab Zaheer. "Machine learning approaches to predict adsorption performance of sugarcane derived-carbon dot− based composite in the removal of dyes." Separation and Purification Technology 351 (2024): 127937.

[29] Wang, Shikai, Kangming Xu, and Zhipeng Ling. "Deep learning-based chip power prediction and optimization: An intelligent EDA approach." International Journal of Innovative Research in Computer Science & Technology 12, no. 4 (2024): 77-87.

[30] Ahmed, Shams Forruque, Md Sakib Bin Alam, Maruf Hassan, Mahtabin Rodela Rozbu, Taoseef Ishtiak, Nazifa Rafa, M. Mofijur, A. B. M. Shawkat Ali, and Amir H. Gandomi. "Deep learning modelling techniques: current progress, applications, advantages, and challenges." Artificial Intelligence Review 56, no. 11 (2023): 13521-13617.

[31] Sengupta, Saptarshi, Sanchita Basak, Pallabi Saikia, Sayak Paul, Vasilios Tsalavoutis, Frederick Atiah, Vadlamani Ravi, and Alan Peters. "A review of deep learning with special emphasis on architectures, applications and recent trends." Knowledge-Based Systems 194 (2020): 105596.

[32] Bokka, Karthiek Reddy, Shubhangi Hora, Tanuj Jain, and Monicah Wambugu. Deep Learning for Natural Language Processing: Solve your natural language processing problems with smart deep neural networks. Packt Publishing Ltd, 2019.

[33] Tunstall, Lewis, Leandro Von Werra, and Thomas Wolf. Natural language processing with transformers. " O'Reilly Media, Inc.", 2022.

[34] Sorin, Vera, Yiftach Barash, Eli Konen, and Eyal Klang. "Deep learning for natural language processing in radiology—fundamentals and a systematic review." Journal of the American College of Radiology 17, no. 5 (2020): 639-648.

[35] Mohsan, Syed Agha Hassnain, Nawaf Qasem Hamood Othman, Yanlong Li, Mohammed H. Alsharif, and Muhammad Asghar Khan. "Unmanned aerial vehicles (UAVs): Practical aspects, applications, open challenges, security issues, and future trends." Intelligent Service Robotics 16, no. 1 (2023): 109-137.

[36] Heidari, Arash, Nima Jafari Navimipour, Mehmet Unal, and Guodao Zhang. "Machine learning applications in internet-of-drones: Systematic review, recent deployments, and open issues." ACM Computing Surveys 55, no. 12 (2023): 1-45.

[37] Olayiwola, Olufemi, Miles Elsden, and Mahmoud Dhimish. "Robotics, Artificial Intelligence, and Drones in Solar Photovoltaic Energy Applications—Safe Autonomy Perspective." Safety 10, no. 1 (2024): 32.

[38] Liu, Yalin, Hong-Ning Dai, Qubeijian Wang, Mahendra K. Shukla, and Muhammad Imran. "Unmanned aerial vehicle for internet of everything: Opportunities and challenges." Computer communications 155 (2020): 66-83.

[39] Majumdar, Parijata, Sanjoy Mitra, Diptendu Bhattacharya, and Bharat Bhushan. "Enhancing sustainable 5G powered agriculture 4.0: Summary of low power connectivity, internet of UAV things, AI solutions and research trends." Multimedia Tools and Applications (2024): 1-45.

[40] Esenogho, Ebenezer, Karim Djouani, and Anish M. Kurien. "Integrating artificial intelligence Internet of Things and 5G for next-generation smartgrid: A survey of trends challenges and prospect." Ieee Access 10 (2022): 4794-4831.

[41] Franki, Vladimir, Darin Majnarić, and Alfredo Višković. "A comprehensive review of Artificial Intelligence (AI) companies in the power sector." Energies 16, no. 3 (2023): 1077.

[42] Heidari, Arash, Nima Jafari Navimipour, Mehmet Unal, and Guodao Zhang. "Machine learning applications in internet-of-drones: Systematic review, recent deployments, and open issues." ACM Computing Surveys 55, no. 12 (2023): 1-45.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 448-456. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3035

RESEARCH ARTICLE OPEN ACCESS

ERGONOMICS THERMAL ENGINEERING APPROACH IN DESIGNING TRADITIONAL DRYER WITH RICE HUSK ENERGY FOR SMALL-SCALE *RENGGINANG* PRODUCTION

I Gede Bawa Susana[1], and Ida Bagus Alit[2]

[1, 2]Department of Mechanical Engineering and Industrial, Faculty of Engineering, University of Mataram, Jl. Majapahit No. 62 Mataram-Nusa Tenggara Barat 83125, Indonesia

[1]https://orcid.org/0000-0003-1597-154X, [2]https://orcid.org/0000-0002-2444-8972

Email: gedebawa@unram.ac.id, alit.ib@unram.ac.id

ARTICLE INFO

Article History
Received: December 1, 2025
Revised: December 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Rengginang,
Rice husk,
Thermal ergonomics,
Traditional drying.

ABSTRACT

Drying is useful for extending the shelf life of a food product by preventing spoilage, making it a crucial step in the *Rengginang* production process to reduce the water content to a safe level. Traditional drying methods, which rely on sunlight, have several limitations, such as dependence on weather, long processing times, and the risk of contamination. This study aims to evaluate the effectiveness of drying rice crackers using thermal energy from rice husks from a technical and ergonomic perspective. The study was conducted to design a dryer using worker anthropometric data and two 360-minute drying scenarios: one using a rice husk-based dryer and one using direct solar drying. The results show that the height of the top and bottom shelves in the drying room is 1285.39 mm and 617.64 mm, based on anthropometric data of standing eye height and gripping hand height. The maximum height of the control device and the height of the door handle are 1649.01 mm and 876.98 mm, based on anthropometric data of standing hand reach and elbow height. The dimensions of the shelves for placing *Rengginang* have a length and width of 711.96 mm and 525.49 mm, respectively, based on anthropometric data of elbow reach and front reach of the gripping hand. The use of a rice husk energy dryer resulted in an average drying temperature of 69.06°C with a range of 51.29-81.90°C, and the initial moisture content of *Rengginang* of 48.27% was successfully reduced to 8.78%. When dried directly in the sun, the moisture content only dropped to 29.21%. This difference indicates a significant influence of technical conditions on drying performance, such as heat distribution and air ventilation. From an ergonomics perspective, drying efficiency contributes to reduced work duration, heat exposure, and worker physical strain. Optimal drying can reduce the risk of musculoskeletal disorders and improve work comfort. Ergonomics thermal engineering results in effective utilization of rice husk energy, including technical aspects and the design of tools and work processes that are appropriate for worker posture.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Traditional food, as a small business, plays a vital role in supporting the rural economy in Indonesia, particularly in Lombok. One popular product is *Rengginang*, a traditional snack made from sticky rice that requires a drying process as a crucial step in its production. In general, the drying process for *Rengginang* is still traditionally carried out by drying it directly in the sun. This method is highly dependent on weather conditions and takes a long time. It also carries the risk of contamination and inconsistent product quality. Direct sun drying carried out in open areas causes losses such as insect infection, uncontrolled and long drying, dust and dirt contamination, and food is susceptible to microbiological contamination by fungi and pathogens as a result of unhygienic conditions in open environments [1], [2].

Rengginang, as a wet food, is also susceptible to contamination by dust, dirt, and microbiological agents such as mold if left in the open for too long. *Rengginang* are traditional rice crackers made from coarsely ground glutinous rice, dried, and then fried until crispy. Although Lombok is located in a tropical region with almost constant sunshine, the temperature is not optimal for the drying process. This is indicated by the long drying time, which automatically stops when the weather is cloudy or rainy. Direct sunlight results in an average drying temperature of 29.44°C [3]. On the other hand, agricultural waste such as rice husks has significant potential as a renewable alternative energy source.

II. THEORETICAL REFERENCE

Lombok Island is an agricultural region where most farmers grow rice as a staple crop. For example, in 2024, according to data from the Central Statistics Agency (BPS), rice production reached 785,918 tons of dry milled grain on Lombok Island, one of Indonesia's agricultural regions [4]. The by-product of processing rice is 20% in the form of rice husks [5], [6]. Theoretically, the potential for rice husks on Lombok Island for 2024 is 157,184 tons [7]. Improper processing of rice husks, which are the main by-product of rice production, causes environmental impacts such as solid waste [8]. The calorific value of rice husk from several research results which have a calorific value equivalent to half the calorific value of coal includes11-15,3 MJ/kg [9], net calorific value 12-16 MJ/kg [10], 13-19 MJ/kg with an average of 18 MJ/kg [11], heat heating value 18.14 ± 0.04 MJ/kg [12]. Rice residue in the form of rice husk and rice straw is used as biofuel to supplement or replace fossil fuels. It has an acceptable calorific value, good thermal characteristics, and low gas emissions in both oxidizing and inert atmospheres, making it a sustainable biomass option for economic and environmental growth [12], [13].

The use of rice husk biomass can be considered as a substitute for fossil fuels, the demand for which is increasing, and the price is increasing, especially for rural communities with small-scale agriculture. Fuel demand is influenced by fuel prices and population growth, which impact economic stability and environmental protection [14]. Rice husks are a biomass waste abundant in agricultural areas, but their use as an energy source for small businesses remains very limited. Proper processing of rice husks can produce an efficient and environmentally friendly drying fuel. Designing a rice husk-based dryer requires consideration of technical and thermal aspects as well as the comfort and safety of business operators. These technical and thermal aspects can be addressed through an energy conversion process using a heat exchanger to convert ambient air into heat. The use of a heat exchanger with rice husk energy can produce a drying temperature between 73 and 119°C, and a mixture of rice husk energy and solar energy produces an average combustion chamber temperature of 347.3°C [15], [16].

Rice husk is suitable for energy use as biomass fuel because of its high silica content, and is superior to rice straw [17]. A heat exchanger can increase the temperature during the drying process by converting biomass energy into thermal energy. A heat exchanger implements the heat transfer process between two fluids at different temperatures separated by a wall without mixing. It can also be used to recover waste heat to reduce the need for additional heating [18-20]. Occupational comfort and safety aspects influence the productivity levels of workers involved in outdoor work due to heat exposure. High heat exposure results in low productivity for outdoor workers, particularly in tropical climates. 60% of workers experience a decrease in productivity when exposed to high heat (WBGT >28°C or temperature >35°C), 30% in indoor or outdoor industries during heat stress conditions, with a decrease in productivity of 2.6% for each degree above the wet bulb temperature of 24°C (WBGT), and a significant economic burden due to work-related heat stress [21-23].

Outdoor workers have a number of risk factors, namely heat stress and heat-related conditions, which affect health, well-being, and productivity, as well as work-related injuries, reduced psychomotor performance, loss of concentration, reduced alertness, loss of coordination, fatigue, and affect productivity and sectoral economies, especially in areas that are highly susceptible to heat [24-27]. Therefore, an ergonomics thermal engineering approach is needed, namely a multidisciplinary approach that integrates the principles of energy efficiency (thermal) and operator comfort and safety (ergonomics) in tool design. Ergonomics assessment is important to increase productivity and reduce risks in physical work [28]. Current ergonomics interventions have had an impact on reducing musculoskeletal disorders (MSDs) complaints widely [29]. The decline in performance that influences the decline in work productivity is the impact of various complaints, both physical and psychological [30].

The Occupational Safety and Health Administration (OSHA) provides recommendations regarding ergonomic measures to prevent sources of disease, namely (1) engineering, such as the design of work tools and work stations; (2) management engineering, such as work organization. These recommendations reduce musculoskeletal disorders in workers to avoid the risk of work injuries [31]. Ergonomics thermal engineering, such as the use of cross ventilation, shade nets, and tool design at optimal heights, has been proven to reduce heat loads and the risk of worker fatigue [32]. Thermal comfort for outdoor applications has more than 20 indices, human psychological states, and other parameters, but most of the models are based on physiological and psychological levels [33].The combination of rice husk-fueled dryer design and worker-oriented design is known as ergonomics thermal engineering. This approach is expected to produce a dryer design that is not only energy efficient and uses renewable fuels but also ergonomic in its use. Thus, the application of appropriate technology based on ergonomic thermal engineering can increase productivity, product quality, and the sustainability of small businesses such as *Rengginang* production at the household level.

III. MATERIALS AND METHODS

The study was conducted using an ergonomics thermal engineering approach. The ergonomics approach was implemented through the participation of workers from small *Rengginang* (rice crackers) businesses who dry the product. The study used worker anthropometric data, a Nordic Body Map (NBM) questionnaire, percentile tables, and materials such as rice husks, *Rengginang* (rice crackers), drying equipment, a drying device with a vertical shelf-shaped drying chamber, and two biomass furnaces. The dryer design was based on user participation, with dimensions based on worker anthropometric data and readily available dryer materials. This ensured the dryer was appropriate for the needs of small *Rengginang* businesses.

The dryer was designed for rice husk energy, and the dryer dimensions were based on workers' anthropometric data. Female workers were first subjectively assessed for musculoskeletal complaints about their work posture during the *Rengginang* drying process.

Musculoskeletal complaints were measured using the Nordic Body Map (NBM) questionnaire. The Nordic Body Map (NBM) measures and identifies muscle and skeletal complaints in various parts of the worker's body due to their work posture. This tool is handy in the initial assessment of ergonomics risks, especially in work environments that require physical activity, such as lifting, bending, prolonged sitting, or repetitive movements. This questionnaire is standard and widely used in ergonomics research due to its low cost, worker involvement, and ease of data collection.

The Nordic Body Map questionnaire was used to identify MSDs symptoms in workers related to ergonomics risks, measuring the risk of injury in human-machine interaction activities [34], [35]. The ergonomic approach was limited to dryer design, while the thermal approach was carried out through direct sunlight drying tests and a drying device. *Rengginang* samples weighing approximately 5 kilograms were used for testing in the drying chamber, and approximately 1 kilogram for testing in direct sunlight. Ergonomics analysis was conducted by comparing them with related research results, while thermal analysis was performed by measuring the ambient and drying chamber temperatures and moisture content. Temperature measurements in the dryer were performed using a K-type thermocouple and a data logger. The dryer design used in this study is a modification and application of Patents IDS000004448 and IDS000005388 [36], [37].

The measurement of workers' anthropometric data was adjusted to the dryer design's needs. The subjects were 25 female workers taken from one of the traditional *Rengginang* businesses in Mataram, Lombok, Indonesia. Measurements include eye height in a standing position, namely for the height of the top shelf, the height of the hand gripping for the height of the bottom shelf, the height of the hand reach in an upright standing position for the maximum height limit of the control device, the height of the standing elbow for the height of the door handle, the width of the elbow span for the maximum length of the tray, and the front reach of the hand gripping for the maximum width of the tray. Anthropometric data were then calculated using percentile values. In this case, the 5th percentile ensures that all subjects can use the dryer design if implemented in the prototype. The 5th percentile is shown in Equation 1.

$$\text{Percentile } 5 = \bar{x} - (1.65\ \delta) \tag{1}$$

The 5th percentile is calculated based on the mean value ($\bar{x}$) and standard deviation (δ) from measurement data. Standard deviation (δ) as in Equation 2, where x_i is the measurement data, $\bar{x}$ is the mean of the measured data, and n is the number of measured data.

$$\delta = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \tag{2}$$

Dryer testing was conducted to dry *Rengginang* by measuring the ambient temperature, drying chamber temperature, and outlet temperature. The testing time was 360 minutes to avoid over-drying and ensure stable drying results. Furthermore, the traditional drying process for *Rengginang* in the sun can take 6–12 hours, depending on the weather, and a 360-minute test is a realistic representation. The calculation of the %wb (wet basis) moisture content of *Rengginang* uses several equations. The wet moisture content (wb) is calculated based on the total weight.

$$\text{Initial Water Mass} = \text{Initial Material Mass x Initial Material Moisture Content} \tag{3}$$

$$\text{Dry Mass} = \text{Initial Material Mass} - \text{Initial Water Mass} \tag{4}$$

$$\text{Final Water Mass} = \text{Final Material Mass} - \text{Dry Mass} \tag{5}$$

$$\text{Final Moisture Content (wb)} = \frac{\text{Final Water Mass}}{\text{Final Material Mass}} \text{ x } 100\% \tag{6}$$

In the research, the results of the dryer design have not been tested on humans or workers, only limited to the use of workers' anthropometric data.

IV. RESULTS AND DISCUSSIONS

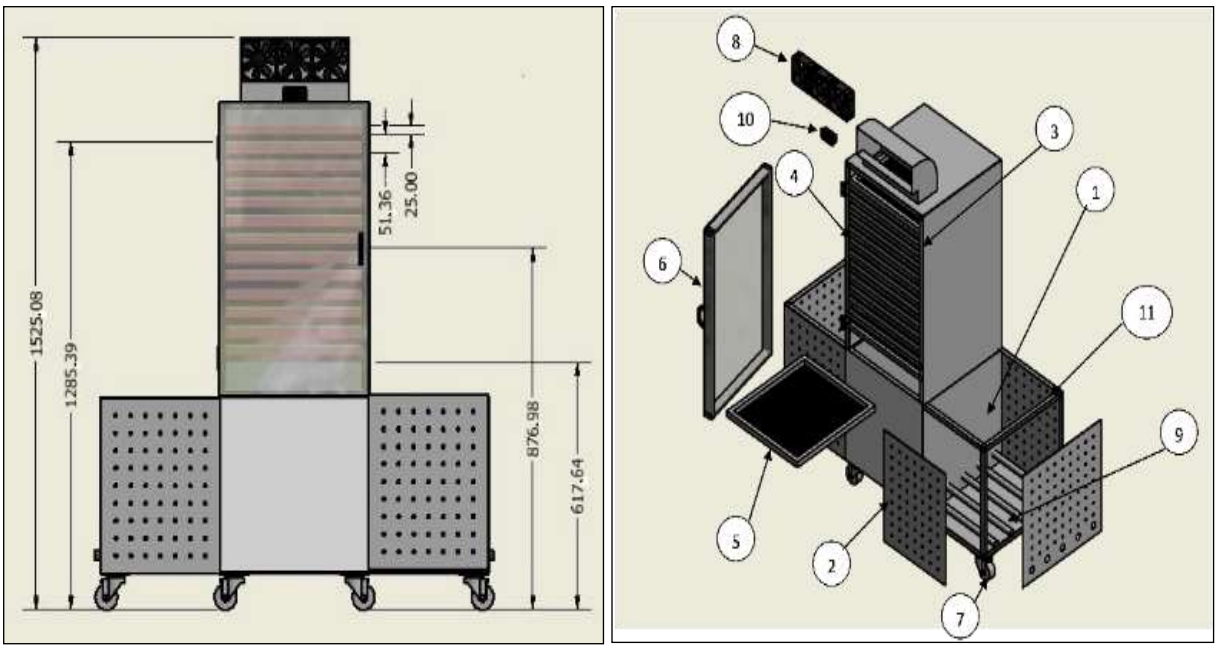
The results of measuring the level of musculoskeletal complaints using the Nordic Body Map questionnaire on several workers who carried out the *Rengginang* drying process found that after work, the average was 55. This total score has a moderate risk level and requires corrective action. The total individual score from the measurement of musculoskeletal complaints using the Nordic Body Map 50-70 has a moderate risk level and requires improvements to the work tools used [30]. Musculoskeletal complaints in workers are caused by unnatural working postures and exposure to the heat of the sun due to work being carried out in open areas during the *Rengginang* drying process, as shown in Figure 1. Figure 1 is based on the guidance of a student's final assignment that has not been published [38]. Figure 1 shows the work process carried out with an unnatural posture.



Figure 1: The worker's posture when drying *Rengginang*.
Source: [38].

Work processes carried out in unnatural postures can result in musculoskeletal complaints with symptoms including pain in the upper and lower back, neck, shoulders, hands, wrists, elbows, knees and ankles, as well as an increased risk of work injuries [31], [39], [40]. An ergonomics approach to small-scale *Rengginang* production was applied through worker participation, namely the use of anthropometric data. The results of anthropometric data measurements from 25 workers obtained an average standing eye height of 137.732 cm with a standard deviation of 5.572 and a 5th percentile value of 128.539 cm. The average hand grip height was 66.888 cm with a standard deviation of 3.105, and the 5th percentile value was 61.764 cm. The average hand reach height was 177.732 cm with a standard deviation of 7.776, and the 5th percentile value was 164.901 cm. The average standing elbow height was 94.86 cm with a standard deviation of 4.341, and the 5th percentile value was 87.698 cm. The average elbow span width was 75.9 cm with a standard deviation of 2.851, and the 5th percentile value was 71.196 cm.

The average front reach of the grasping hand is 57.90 cm with a standard deviation of 3.241, and the 5th percentile value is 52.549 cm. The application of the 5th percentile from the results of anthropometric data measurements is used as a maximum dimension requirement for the dryer design in this study. The height of the top and bottom shelves in the drying chamber is 1285.39 mm and 617.64 mm, based on anthropometric data of standing eye height and grasping hand height. The maximum height of the control device and the height of the door handle are 1649.01 mm and 876.98 mm, based on anthropometric data of the hand reach height and standing elbow height. The dimensions of the rack, which is used to place the *Rengginang*, have a length and width of 711.96 mm and 525.49 mm, respectively, based on anthropometric data of the elbow reach width and the front reach of the hand gripping. These results are applied to the design of the dryer as presented in Figure 2. Based on the design in Figure 2, the application is in the form of a prototype drying tool with a *Rengginang* drying process as shown in Figure 3.



1. Furnace, 2. Air circulation holes, 3. Drying chamber, 4. Drying shelf limit, 5. Tray, 6. Drying chamber door, 7. Wheel, 8. Exhaust fan, 9. Heat exchanger pipe, 10. Thermostat.

Figure 2: Dryer design with dimensions based on worker anthropometric data, in millimetres (mm).
Source: [38].



(a) (b)

Figure 3: Drying of *Rengginang*: a) rice husk energy dryer, b) direct sunlight.
Source: Authors, (2026).

The drying process for *Rengginang* using a dryer designed with an ergonomics thermal approach, as shown in Figure 3(a), can be carried out indoors. This ensures workers are not exposed to the sun's heat and allows for a more natural working posture, as seen from the use of vertical racks in the drying room. The energy source uses rice husk biomass. The dryer is also equipped with solar cells to drive a fan, ensuring heat transfer occurs through forced convection. This is in line with the research of Duque-Dussán et al., which shows that hybrid solar dryers that combine solar and mechanical drying principles are equipped with biomass burners and photovoltaic systems, allowing for completely independent drying [41]. A shelf adapted to anthropometric data makes it easier for operators to retrieve/insert trays without bending. Furthermore, the resulting product is more hygienic because the drying process takes place in a closed drying room.

Workers no longer have to monitor the dried product, and it doesn't need to be turned over. Ergonomics interventions through engineering have resulted in a significant reduction in ergonomics risk factors and reduced rates of musculoskeletal disorders among workers [42]. Ergonomics approaches are effective in reducing postural loads and can be seen as an opportunity for farmers to increase attention to health and well-being in farming [43]. Figure 3(b) shows the direct sun-drying process carried out in an open space. This leaves the dried product vulnerable to dust and other dirt, and workers exposed to the sun's heat. Furthermore, the drying temperature is suboptimal because it uses ambient temperature. Figure 4 shows the results of temperature measurements for the *Rengginang* drying process using rice husk energy dryers and direct sunlight.

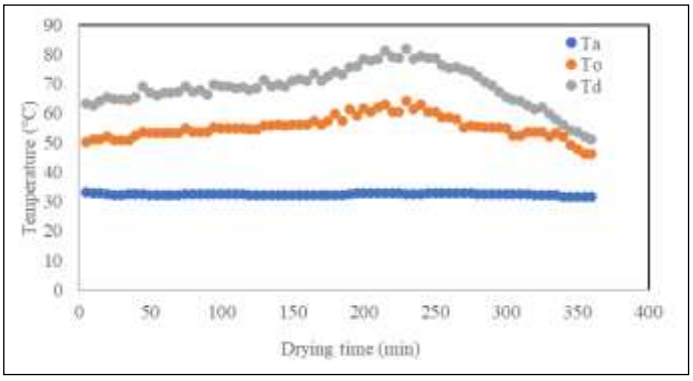


Figure 4: Distribution of drying temperature of *Rengginang* with a dryer and direct sunlight.
Source: Authors, (2026).

The average ambient temperature (Ta) used to dry *Rengginang* was 32.42°C, with a range of 31.54-33.04°C. Temperature measurements were taken from the first 5 minutes to 360 minutes, from 9:00-15:00 WITA. This is in line with Sarmad et al., who found that solar radiation intensity varies from 9:00-15:00 and increases from morning until reaching its maximum value at 13:00 [44]. The use of a dryer powered by rice husk energy resulted in an average drying temperature (Td) of 69.06°C with a range of 51.29-81.90°C. This temperature range resulted from using rice husk in the furnace for only one filling and without adding any additional rice husk mass. Furthermore, the lower initial temperature resulted from the rice husk undergoing a prior evaporation process. The moisture content of rice husk is 6-10%, and there is a drying zone in the biomass, namely the process of evaporating the water content before further heating occurs [45-47]. When rice husks are converted into charcoal, high temperatures are generated. Test results also show that the outlet temperature of the drying chamber (To) remains quite high, averaging 55.29°C and ranging from 46.2 to 64.11°C. The drying temperatures inside and outside the drying chamber were 113.02% and 70.54% higher, respectively, compared to ambient temperature. At these temperatures, Figure 5 shows the decrease in the mass of the dried *Rengginang*.

Figure 5 shows a comparison of the decrease in *Rengginang* mass during the 360-minute test between the rice husk dryer and direct sun drying. The average mass of *Rengginang* in the rice husk energy dryer was 1086 g on each rack, while direct solar drying was 1086 g. The use of the rice husk energy dryer showed a faster decrease in *Rengginang* mass during 360 minutes compared to direct sun drying. This is influenced by the higher temperature used in the rice husk dryer. The rice husk dryer produced a final mass of 616.2 g, while direct sunlight drying produced 794 g. *Rengginang* initial moisture content is 48.27% wb. The moisture content %wb (wet basis) is the percentage of moisture content in a material calculated based on its wet weight.

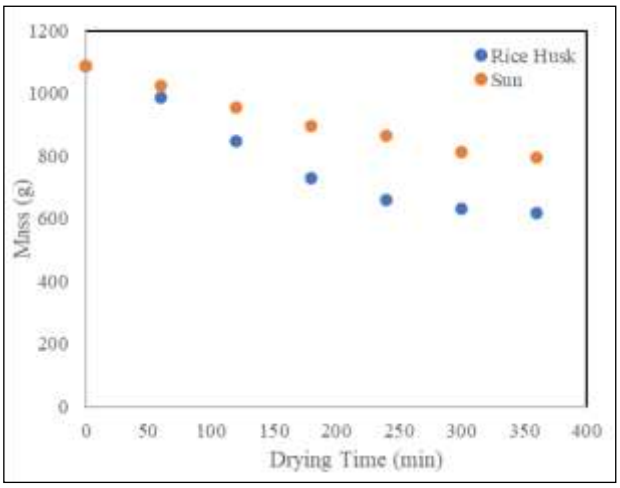


Figure 5: Comparison of drying *Rengginang* using rice husk dryer and direct sunlight.
Source: Authors, (2026).

In this study, drying *Rengginang* for 360 minutes showed that drying using rice husk energy accelerated the reduction in the product's water content compared to conventional solar drying. Comparison data for drying using rice husk energy and direct sun drying are presented in Table 1.

Table 1: Comparison of data on rice husk energy dryer and direct sun drying.

Parameter	Rice husk energy dryer	Direct sun drying
Mass of initial materials (g)	1086	1086
Initial moisture content of material (%)	48.27	48.27
Final material mass (g)	616.2	794
Final moisture content (wb) (%)	8.83	29.25
Drying time (minutes)	360	360

Source: Authors, (2026).

Based on Table 1, the rice husk energy dryer significantly reduced the initial moisture content of *Rengginang* by 48.27% to 8.83% in 360 minutes. This value is within the ideal moisture content limit for *Rengginang* products ready for frying (<12%). The water content of *Rengginang* according to SNI 01-4307-1996 regarding rice crackers is a maximum of 12% [48]. This demonstrates that the rice husk energy dryer is operating optimally, with stable and efficient heat distribution and airflow. Meanwhile, sun direct drying yielded a final moisture content of 29.25% within the same drying time. This indicates that the drying process is not optimal due to differences in heating temperature and uneven air circulation. Table 1 shows how drying performance is highly dependent on technical conditions such as temperature, airflow, and heat distribution. Rice husk energy dryers provide stable heat and can be used at any time (independent of the weather), thus offering advantages in terms of time and quality. In addition, rice husks' abundant availability can reduce waste through their use in drying energy. Vast expanses of rice fields have the potential to utilize rice husk waste as a biomass raw material to support renewable energy [49]. Meanwhile, sun direct drying often requires more than a full day to reduce the moisture content to less than 12% due to its high dependence on weather, humidity, and direct sunlight.

Ergonomics studies have shown that these differences in performance have direct implications for workers' physical workload and duration of heat exposure. More energy-efficient rice husk drying means shorter labour time to achieve the final result, thus reducing exposure to high temperatures and lowering the risk of muscle fatigue and heat stress. This is particularly relevant in traditional food processing activities such as *Rengginang*, which are generally carried out manually in high-temperature environments. Conversely, sun direct drying shows suboptimal drying, which extends the work duration and increases static loads due to the need to regulate and monitor the dryer for longer periods. Furthermore, this inefficiency can impact the frequency of stirring and tray rearrangement activities, as well as increase muscle fatigue in the arms, back, and lower legs due to unnatural working postures. Therefore, an ergonomics thermal engineering approach in designing traditional rice husk energy dryers is not only technically superior but can also provide significant ergonomics benefits, such as reducing prolonged heat exposure, shortening work cycles, reducing the risk of musculoskeletal disorders, and increasing work comfort. This application, which is centered on humans as workers, is in line with research results that state humans remain important for adaptability and innovation, so human-machine integration is a crucial factor [50].

The importance of an ergonomics approach in evaluating the performance of drying equipment so that technological innovation prioritizes process efficiency and the long-term safety and health of workers.

V. CONCLUSIONS

Drying *Rengginang* using rice husk energy for 360 minutes can increase the ambient temperature and optimally reduce the moisture content. The initial moisture content of 48.27% can be reduced to 8.83% when using rice husk energy, while under direct sunlight conditions, the reduction only reaches 29.25%. This indicates that factors such as drying temperature, air circulation, and rice husks influence the drying duration. From an ergonomics perspective, drying time impacts not only product quality but also workers' physical workload. Fast and effective drying can reduce heat exposure, repetitive manual activities, and the risk of musculoskeletal disorders. Conversely, suboptimal drying can potentially extend work time and increase physical fatigue. Utilizing rice husk energy as a heat source not only offers promising efficiency and sustainability but also contributes to increased work comfort if supported by ergonomic tools and process design. The practical implications of these results encourage the need to integrate technical and ergonomics approaches in the development of small and medium-scale food drying systems.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: I Gede Bawa Susana and Ida Bagus Alit.
Methodology: I Gede Bawa Susana.
Investigation: I Gede Bawa Susana and Ida Bagus Alit.
Discussion of results: I Gede Bawa Susana and Ida Bagus Alit.
Writing – Original Draft: I Gede Bawa Susana.
Writing – Review and Editing: I Gede Bawa Susana and Ida Bagus Alit.
Resources: I Gede Bawa Susana and Ida Bagus Alit.
Supervision: I Gede Bawa Susana and Ida Bagus Alit.
Approval of the final text: I Gede Bawa Susana and Ida Bagus Alit.

VII. ACKNOWLEDGMENTS

The author would like to thank the Department of Mechanical Engineering, Faculty of Engineering, University of Mataram for facilitating the implementation of this research, and also to I Wayan Gandhi Artha for his participation in collecting anthropometric data on *Rengginang* workers.

VIII. REFERENCES

[1] J. Petikirige, A. Karim, and G. Millar, "Effect of Drying Techniques on Quality and Sensory Properties of Tropical Fruits", International Journal of Food Science and Technology, vol. 57, 2022, pp. 6963–6979. https://doi.org/10.1111/ijfs.16043.

[2] D. Alp, and Ö. Bulantekin, "The Microbiological Quality of Various Foods Dried by Applying Different Drying Methods: A Review", European Food Research and Technology, vol. 247, 2021, pp. 1333–1343. https://doi.org/10.1007/s00217-021-03731-z.

[3] I G. Bawa Susana, I.B. Alit, and I D.K. Okariawan, "Rice Husk Energy Rotary Dryer Experiment for Improved Solar Drying Thermal Performance on Cherry Coffee", Case Studies in Thermal Engineering, vol. 41, 2023, pp. 102616. https://doi.org/10.1016/j.csite.2022.102616.

[4] BPS., Berita Resmi Statistik, "Luas Panen dan Produksi Padi di Provinsi Nusa Tenggara Barat, 2024", Available online: https://ntb.bps.go.id/en/statistics-table/3/WmpaNk1YbGFjR0pOUjBKYWFIQlBSU3MwVHpOVWR6MDkjMw==/luas-panen--produktivitas--dan-produksi-padi-menurut-kabupaten-kota-di-provinsi-nusa-tenggara-barat--2024.html (accessed on March 17, 2025).

[5] S.K.S. Hossain, L. Mathurand, and P.K. Roy, "Rice Husk/Rice Husk Ash As An Alternative Source of Silica in Ceramics: A Review", Journal of Asian Ceramic Societies, vol. 6, no. 4, 2018, pp. 299–313. https://doi.org/10.1080/21870764.2018.1539210.

[6] Dinas Pertanian Pemerintah Kabupaten Buleleng, "Pembuatan Arang Sekam Padi Sebagai Media Tanam di BPP Seririt", Available online: https://distan.bulelengkab.go.id/informasi/detail/berita/67_pembuatan-arang-sekam-padi-sebagai-media-tanam-di-bpp-seririt (accessed on March 2, 2025).

[7] I G. Bawa Susana, I K.P. Putra, and I G.A.K.C.A.W. Aryadi, "Analysis of Rice Husk Biomass Potential in Lombok Island, Indonesia as A Renewable Energy Source", Energy, Materials and Product Design, vol. 4, no. 1, 2025, pp. 260-265. https://doi.org/10.29303/empd.v4i1.6877.

[8] S.D.S. Piyathissa, P.D. Kahandage, Namgay, H. Zhang, R. Noguchi, and T. Ahamed, "Introducinga Novel Rice Husk Combustion Technology for Maximizing Energy and Amorphous Silica Production Using a Prototype Hybrid Rice Husk Burner to Minimize Environmental Impacts and Health Risk", Energies, vol. 16, no. 3, 2023, pp. 1120. https://doi.org/10.3390/en16031120.

[9] J.O. Awulu, P.A. Omale, and J.A. Ameh, "Comparative Analysis of Calorific Values of Selected Agricultural Wastes," Nigerian Journal of Technology (NIJOTECH), vol. 37, no. 4, 2018, pp. 1141-1146. https://doi.org/10.4314/njt.374.1881.

[10] International Finance Corporation, "Converting Biomass to Energy: A Guide for Developers and Investors", N.W. Washington, D.C.: Pennsylvania Avenue, June 2017.

[11] J. Smith, "Combined Heat and Power from Rice Husks", England, London: GMB Energy Central, 2007.

[12] J.I. Arranz, M.T. Miranda, I. Montero, and F.J. Sepúlveda, "Thermal Study and Emission Characteristics of Rice Husk Using TG-MS", Materials, vol. 14, no. 20, 2021, pp. 6203. https://doi.org/10.3390/ma14206203.

[13] S. Kaniapan, J. Pasupuleti, K. Patma Nesan, H.N. Abubackar, H.A. Umar, T.L. Oladosu, S.R. Bello, and E.R. Rene, "A Review of the Sustainable Utilization of Rice Residues for Bioenergy Conversion Using Different Valorization Techniques, Their Challenges, and Techno-Economic Assessment", International Journal of Environmental Research and Public Health, vol. 19, no. 6, 2022, pp. 3427. https://doi.org/10.3390/ijerph19063427.

[14] M.A. Gharaibeh, R. Al-Jarrah, and A. Alkhatatbeh, “A Study on the Fuel Consumption Demands of the Jordanian Residential Sector”, Jordan Journal of Mechanical and Industrial Engineering, vol. 18, no. 1, 2024, pp. 191-205. https://doi.org/10.59038/jjmie/180115.

[15] I.B. Alit, I.G. Bawa Susana, and I.M. Mara, “Thermal Study of the Dryer with Heat Exchanger Pipe Installed in Rice Husk Double Furnace During Drying White Turmeric”, Frontiers in Heat and Mass Transfer, vol. 19, no. 1, 2022, pp. 1-7. https://doi.org/10.5098/hmt.19.25.

[16] Y.C. Mihret, M.A. Delele, and S.T. Hailemesikel, “Design, Development, and Testing of Rice-Husk Fueled Mixed-Flow Rice Dryer for Small-Scale Rice Producer Farmers”, Heliyon, vol. 9, no. 7, 2023, pp. e18077. https://doi.org/10.1016/j.heliyon.2023.e18077.

[17] C.-H. Tsai, Y.-H. Shen, and W.-T. Tsai, “Thermochemical Characterization of Rice-Derived Residues for Fuel Use and Its Potential for Slagging Tendency”, Fire, vol. 6, no. 6, 2023, pp. 230. https://doi.org/10.3390/fire6060230.

[18] T.L. Bergman, A.S. Lavine, F.P. Incropera, and D.P. DeWitt, “Fundamentals of Heat and Mass Transfer”, 7th ed., New York: John Wiley & Sons, 2011.

[19] Y.A. Çengel, “Heat Transfer: A Practical Approach”, 2nd ed., New York: McGraw-Hill, 2002.

[20] Y.S.M. Goselink, and B.C. Ramirez, “Characterization of an Air-to-Air Heat Exchanger for Manure Belt Drying Ventilation in an Aviary Laying Hen House”, Journal of Applied Poultry Research, vol. 28, no. 4, 2019, pp. 1359-1369. https://doi.org/10.3382/japr/pfz075.

[21] S. Han, L. Dong, Y. Weng, and J. Xiang, “Heat Exposure and Productivity Loss Among Construction Workers: A Meta-Analysis”, BMC Public Health, vol. 24, 2024, pp. 3252. https://doi.org/10.1186/s12889-024-20744-x.

[22] A.D. Flouris, P.C. Dinas, L.G. Ioannou, L. Nybo, G. Havenith, G.P. Kenny, and T. Kjellstrom, “Workers’ Health and Productivity Under Occupational Heat Strain: A Systematic Review and Meta-Analysis”, The Lancet Planetary Health, vol. 2, no. 12, 2018, pp. e521-e531. https://doi.org/10.1016/S2542-5196(18)30237-7.

[23] M.A. Borg, J. Xiang, O. Anikeeva, D. Pisaniello, A. Hansen, K. Zander, K. Dear, M.R. Sim, and P. Bi, “Occupational Heat Stress and Economic Burden: A Review of Global Evidence”, Environmental Research, vol. 195, 2021, pp. 110781. https://doi.org/10.1016/j.envres.2021.110781.

[24] M.A. Idris, C. Markham, K.D. Mena, and W.B. Perkison, “Examining Management and Employees’ Perceptions of Occupational Heat Exposure and the Effectiveness of a Heat Stress Prevention Intervention on Safety and Well-Being among Natural Gas Construction Workers: A Qualitative Field-Based Study”, International Journal of Environmental Research and Public Health, vol. 21, no. 9, 2024, pp. 1255. https://doi.org/10.3390/ijerph21091255.

[25] B.M. Varghese, A. Hansen, Peng Bi, and D. Pisaniello, “Are Workers at Risk of Occupational Injuries Due to Heat Exposure? A Comprehensive Literature Review”, Safety Science, Vol. 110, Part A, 2018, pp. 380-392. https://doi.org/10.1016/j.ssci.2018.04.027.

[26] M. Amoadu, E.W. Ansah, J.O. Sarfo, and T. Hormenu, “Impact of Climate Change and Heat Stress on Workers’ Health and Productivity: A Scoping Review”, The Journal of Climate Change and Health, vol. 12, 2023, pp. 100249. https://doi.org/10.1016/j.joclim.2023.100249.

[27] Global Heat Health Information Network, “Occupational Heat Stress Can Impact the Productivity, Safety and Health of Both Outdoor and Indoor Workers Who Don’t Have Access to Appropriate Cooling”, Available online: https://ghhin.org/at-work/ (accessed on March 8, 2025).

[28] I G. Bawa Susana, I.B. Alit, and I G.A.K.C.A.W. Aryadi, “Aplikasi Ergonomi Berdasarkan Data Antropometri Pekerja pada Desain Alat Kerja”, Energy, Materials and Product Design, vol. 1, no. 1, 2022, pp. 28-34.

[29] L. Benos, D. Tsaopoulos, and D. Bochtis, “A Review on Ergonomics in Agriculture. Part I: Manual Operations”, Applied Sciences, vol. 10, no. 6, 2020, pp. 1905.

[30] Tarwaka, “Ergonomi Industri: Dasar-dasar Pengetahuan Ergonomi dan Aplikasi di Tempat Kerja”, edisi 2 dengan revisi, Surakarta: Harapan Press, 2019.

[31] OSHA, Ergonomics, Available online: https://www.osha.gov/ergonomics (accessed on April 1, 2025).

[32] L.G. Ioannou, K. Mantzios, L. Tsoutsoubi, E. Nintou, M. Vliora, P. Gkiata, C.N. Dallas, G. Gkikas, G. Agaliotis, K. Sfakianakis, A.K. Kapnia, D.J. Testa, T. Amorim, P.C. Dinas, T.S. Mayor, C. Gao, L. Nybo, and A.D. Flouris, “Occupational Heat Stress: Multi-Country Observations and Interventions”, International Journal of Environmental Research and Public Health, vol. 18, no. 12, 2021, pp. 6303. https://doi.org/10.3390/ijerph18126303.

[33] Q. Zhao, Z. Lian, and D. Lai, “Thermal Comfort Models and their Developments: A Review”, Energy and Built Environment, vol. 2, no. 1, 2021, pp. 21-33. https://doi.org/10.1016/j.enbenv.2020.05.007.

[34] A. Kusumawardhani, H. Djamalus, and K.D. Lestari, “Ergonomic Risk Assessment and MSDs Symptoms Among Laboratory Workers Using SNI 9011-2021”, The Indonesian Journal of Occupational Safety and Health, vol. 12, no. SI1, 2023, pp. 35-41. http://doi.org10.20473/ijosh.v12iSI1.2023.35-41.

[35] Y.P. Yusdystira, and Iyan Andriana, “Recommendations for Work Posture Using Nordic Body Maps and RULA Methods”, AIP Conf. Proc., vol. 2882, 2023, pp. 050001. https://doi.org/10.1063/5.0176488.

[36] I.G. Bawa Susana, I.B. Alit, and I.G.N.K. Yudhyadi, “Alat Pengering Berbahan Bakar Limbah Sekam Padi dengan Mekanisme Penukar Kalor”, Kementerian Hukum dan Hak Asasi Manusia Republik Indonesia, no. Paten IDS000004448, 2021.

[37] I.B. Alit, I.G. Bawa Susana, and I.M. Mara, “Pengering Hybrid Surya-Biomassa”, Kementerian Hukum dan Hak Asasi Manusia Republik Indonesia, no. Paten IDS000005388, 2022.

[38] I.W.G. Artha, “Desain Ergonomi Pengering Double Tungku Berbahan Bakar Sekam Padi Berdasarkan Data Antropometri Pekerja pada Pengeringan Rengginang”, Tugas Akhir (Tidak Dipublikasikan), Mataram: Jurusan Teknik Mesin Fakultas Teknik Universitas Mataram, 2023.

[39] R.S. Bridger, “Introduction to the Ergonomics”, 2nd ed., New York: Taylor & Francis, 2003.

[40] T.S. Olowogbon, R.O. Babatunde, E. Asiedu, and A.M. Yoder, “Prevalence and Exposure to Ergonomic Risk Factors Among Crop Farmers in Nigeria”, Applied Sciences, vol. 11, no. 24, 2021, pp. 11989. https://doi.org/10.3390/app112411989.

[41] E. Duque-Dussán, J.R. Sanz-Uribe, and J. Banout, “Design and Evaluation of a Hybrid Solar Dryer for Postharvesting Processing of Parchment Coffee”, Renewable Energy, vol. 215, 2023, pp. 118961. https://doi.org/10.1016/j.renene.2023.118961.

[42] K. Hemati, Z. Darbandi, E. Kabir-Mokamelkhah, M. Poursadeghiyan, M.S. Ghasemi, M. Mohseni-Ezhiye, Y. Abdolahian, M. Aghilinejad, M. Ali Salehi, and N. Dehghan, "Ergonomic Intervention to Reduce Musculoskeletal Disorders Among Flour Factory Workers", Work, Vol. 67, No. 3, 2020, pp. 611-618. https://doi.org/10.3233/WOR-203275.

[43] D. Kee, "Participatory Ergonomic Interventions for Improving Agricultural Work Environment: A Case Study in a Farming Organization of Korea", Applied Sciences, vol. 12, no. 4, 2022, pp. 2263. https://doi.org/10.3390/app12042263.

[44] A. Sarmad, A. Hussein, and N.A. Jabbar, "Investigation of the Effect of Changing Water Heights on the Performance of a Solar Distillation System Using Crushed Porous Coal Rocks", Jordan Journal of Mechanical and Industrial Engineering, vol. 18, no. 1, 2024, pp. 171-178. https://doi.org/10.59038/jjmie/180113.

[45] S. Kamari, and F. Ghorbani, "Extraction of Highly Pure Silica from Rice Husk As An Agricultural By-Product and Its Application in the Production of Magnetic Mesoporous Silica MCM–41", Biomass Conversion and Biorefinery, 2020, pp. 1-9. https://doi.org/10.1007/s13399-020-00637-w.

[46] S. Herodian, "Peluang dan Tantangan Industri Berbasis Hasil Samping Pengolahan Padi", Jurnal Pangan, vol. 16, no. 1, 2007, pp. 38-49.

[47] F. Mufid, and S. Anis, "Pengaruh Jenis dan Ukuran Biomassa Terhadap Proses Gasifikasi Menggunakan Downdraft Gasifier, Rekayasa Mesin, vol. 10, no. 3, 2019, pp. 217-226. https://doi.org/10.21776/ub.jrm.2019.010.03.

[48] Istiqomaturrosyidah, and E.S. Murtini, "Inovation of Rengginang as Food Source of Fiber with The Addition of Undaria pinnatifida Seaweed", Pro Food (Jurnal Ilmu dan Teknologi Pangan), vol. 7, no. 1, 2021, pp. 812-820.

[49] Samsurizal, R.F. Azlinando, and A.N. Afandi, "Techno-Economic Evaluation of Rice Husk Co-Firing as A Sustainable Biomass Fuel Alternative", Journal of Engineering and Technology for Industrial Applications, vol. 10, no. 50, 2024, pp. 28-33. https://doi.org/10.5935/jetia.v10i50.1292.

[50] N.T. Machado, and C.M.T. Rodriguez, "Logistics 5.0 Maturity Model: A Human-Centric and Sustainable Approach for the Supply Chain of the Future", Journal of Engineering and Technology for Industrial Applications, vol. 11, no. 51, 2025, pp. 164-170. https://doi.org/10.5935/jetia.v11i51.1407.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 457-464. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3042

RESEARCH ARTICLE OPEN ACCESS

APPLICATION OF THE ANALYTIC HIERARCHY PROCESS (AHP) METHOD FOR CHOOSING A PORTABLE RETINOGRAPH IN OPHTHALMOLOGICAL CLINICAL PRACTICE

Marcelo Maia do Nascimento[1] and Railma Lima de Paula[2]

[1,2]Post Graduate Program in Electrical Engineering the of Institute of Technology from Federal University of Pará (PPGEE/ITEC/UFPA). Avenue Augusto Correa, Campus: Guamá. Bairro: Guamá. CEP: 66075. Belém - Pará - Brasil. Telephone: (91)32017112.

[1]https://orcid.org/0000-0001-7183-8642, [2]https://orcid.org/0009-0005-9076-8850,

Email: *marcelo.maia.nascimento@gmail.com, railma.lima.paula@gmail.com

ARTICLE INFO

Article History

Received: December 2, 2025

Revised: December 20, 2025

Accepted: January 1, 2026

Published: January 31, 2026

Keywords:

AHP; multi-criteria decision making; portable retinograph; medical technology; equipment selection.

ABSTRACT

The selection of portable retinographs has become a significant challenge in teleophthalmology and clinical services that depend on fast and accurate capture of fundus images. The wide variety of models available on the market, each with distinct technical and operational characteristics, requires a structured evaluation based on multiple criteria. This study aimed to apply the Analytic Hierarchy Process (AHP) method to support multi-criteria decision-making in choosing the portable retinograph best suited to clinical practice needs. Criteria such as image quality, portability, battery life, cost, technical support, and compatibility with digital systems were defined. Based on the hierarchical structuring of the problem and the elaboration of paired comparison matrices, the relative weights of each criterion and the performance of the evaluated alternatives were calculated. The results objectively revealed the most appropriate model according to the priorities established by the experts. It was concluded that AHP proved to be effective, transparent, and replicable, making it a valuable tool for decision-making processes involving medical technology.

Copyright ©2025 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Fundus photography is an essential procedure for detecting and monitoring ocular pathologies such as diabetic retinopathy, glaucoma, vascular occlusions, and macular diseases [1]. With the expansion of teleophthalmology programs and primary care screening, portable fundus cameras have come to play a central role, enabling examinations outside hospital environments and contributing to early diagnosis in remote populations or those with limited access to specialists [2]. The growth of the portable fundus camera market has brought a variety of models differing in cost, image resolution, battery life, ergonomics, and integration with information systems [3]. This diversity makes the selection process complex, especially when budget constraints and specific operational needs are involved. Decisions based solely on the lowest price or isolated features may compromise exam quality and the effectiveness of the service [4]. In this context, multicriteria decision-making (MCDA) methods emerge as alternatives capable of systematizing judgments and balancing conflicting criteria [5]. Among them, the Analytic Hierarchy Process (AHP), developed by Thomas Saaty, stands out for its ability to organize complex problems into hierarchical structures and transform subjective judgments into consistent mathematical values [6]. The objective of this study is to apply the AHP method to select the portable fundus camera most suitable for the needs of an ophthalmology clinic or telemedicine program. To achieve this, relevant decision criteria were defined, the problem hierarchy was constructed, and pairwise comparisons between criteria and alternatives were carried out, allowing the calculation of relative weights and the identification of the best option. The specific objectives include identifying the decision criteria, structuring the hierarchical model of the problem, applying AHP to the available alternatives, and assessing the consistency of the judgments made.

II. RELATED WORK

The application of multicriteria decision-making methods to the selection and management of health technologies has been widely discussed in the literature, particularly in the context of acquiring medical equipment [7]. Within the use of Multicriteria Decision Analysis (MCDA) for medical equipment management, the AHP stands out as one of the most suitable methods to support evaluation, selection, and investment-prioritization tasks in resource-constrained environments [8]. In this scenario, AHP has become established as a tool capable of reducing uncertainties and providing greater transparency to decisions involving multiple technical, economic, and operational criteria [9],[10]. In more specific studies, various authors have proposed AHP-based models to directly support the selection of medical devices. One application of AHP elicited user needs for acquiring a computed tomography scanner, demonstrating that the method enables the integration of preferences from different stakeholders into a structured model for the procurement of diagnostic imaging equipment [11]. Along similar lines, [12] developed a multicriteria methodology based on AHP to select the most ergonomically suitable ultrasound device, considering criteria related to technical performance and the well-being of healthcare professionals.

Other works reinforce the use of AHP as the core of decision-support tools for hospital procurement. A model for weighting criteria for selecting medical equipment based on AHP highlighted its ability to structure and prioritize factors such as cost, maintenance, reliability, and technical support [13],[14]. More recent studies have advanced toward integrated frameworks that combine AHP with other methods—such as TOPSIS, VIKOR, and PROMETHEE—with the aim of increasing the robustness of decisions and incorporating sensitivity-analysis scenarios in the selection of complex medical devices [15]. In the specific domain of imaging technologies and diagnostic services, [16] employed AHP to rank radiology departments, demonstrating the method's potential for capturing aspects of quality, efficiency, and technological availability in the evaluation of diagnostic imaging services. The performance of two portable retinal cameras—one smartphone-based and the other smartscope-based—was assessed for capturing fundus images, highlighting the growing role of portable devices in ophthalmologic practice and their integration with OCT examinations [17].

A recent review addressed the advancement of portable devices in ophthalmology and their implications for screening diseases such as diabetic retinopathy and age-related macular degeneration [18]. Although these studies do not directly apply AHP to the selection of portable fundus cameras, they highlight the growing relevance of these technologies and the need for structured evaluation methods [19]. In light of these works, the present study contributes by applying AHP specifically to the selection of portable fundus cameras, articulating technical criteria (image quality, digital compatibility), operational criteria (portability, battery life), and organizational criteria (cost, technical support). Unlike broader approaches to selecting medical equipment or other imaging devices, this article focuses on the specific context of teleophthalmology and clinical ophthalmic practice, filling a gap in the literature by proposing and testing an AHP model for the comparative selection of different commercial portable fundus-camera models.

III. LITERATURE REVIEW

III.1 ANALYTIC HIERARCHY PROCESS (AHP) METHOD

The Analytic Hierarchy Process (AHP) was developed by Professor Thomas L. Saaty in the early 1970s and later consolidated in a classical work dedicated to the method [20]. It emerged as a response to the need to structure and quantify subjective judgments in complex decision-making processes in which different qualitative and quantitative criteria must be considered jointly and in a comparable manner [21]. Since then, it has become one of the most widely used multicriteria decision-making methods, applied in fields such as healthcare, engineering, logistics, public management, and technology selection [10]. AHP is based on three fundamental principles: decomposition, comparative judgments, and priority synthesis [22]. Initially, the problem is decomposed into a hierarchical structure that facilitates understanding of the decision process. The upper level contains the overall objective; the intermediate levels contain the criteria and subcriteria; and at the base of the hierarchy are the alternatives. To illustrate this organization, it is recommended to include a figure representing the hierarchical structure of the problem, such as a vertical diagram containing the objective, criteria, and alternatives [23], as shown in Figure 1.

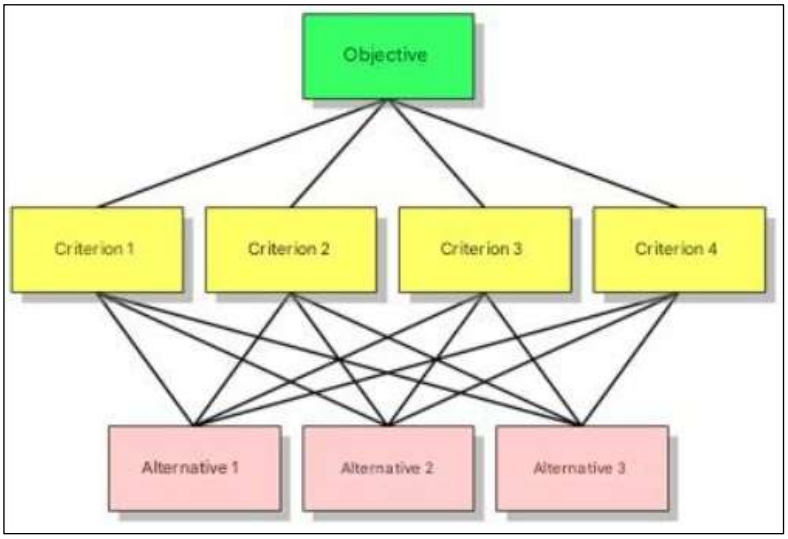


Figure 1: Hierarchical structure of the AHP.
Source: [6].

After the problem has been structured, the stage of comparative judgments begins. The elements at each level of the hierarchy are compared pairwise using Saaty's Fundamental Scale, which ranges from 1 (equal importance) to 9 (extreme importance), including their reciprocal values. This scale enables the intuitive capture of the relative preference between two elements. To reinforce this understanding, it is recommended to include a Table 1 with Saaty's Fundamental Scale, containing descriptions and examples.

Table 1: Saaty scale used in the pairwise comparisons.

	Element 1	**Element 2**	**…**	**Element D**
Element 1	1	a_{12}	…	a_1D
Element 2	$1/a_{21}$	1	…	a_2D
⋮	⋮	⋮	⋮	⋮
Element D	$1/aD_1$	$1/aD_2$	…	1

Source: [6].

The comparisons are organized into square matrices, from which the priority vector is extracted, typically through the principal eigenvector associated with the largest eigenvalue λmax. This vector mathematically represents the weight of each element in relation to the objective [24] ,[25]. An essential aspect of the method is the verification of the consistency of the judgments made. Because human comparisons may naturally present inconsistencies, a specific procedure was established to assess the quality of these comparisons [26]. This process involves calculating the Consistency Index (CI) and the Consistency Ratio (CR), which relates the CI to a reference value known as the Random Index (RI). When the CR is equal to or less than 0.10, the judgments are considered to have acceptable coherence, indicating that the comparisons follow a logically consistent pattern [27]. Another important contribution of AHP is its ability to integrate judgments from multiple experts. When more than one evaluator is involved, judgments can be aggregated using the geometric mean, preserving independence and valuing collective expertise.

This step is particularly relevant in technological decision-making in healthcare, where shifts in priorities can significantly impact the final choice [28]. Several studies demonstrate the versatility and potential of AHP when applied to complex technological systems. Different authors have used the method in areas such as software selection, educational platform choice, and supplier evaluation in the healthcare sector, highlighting its ability to handle multiple—and often conflicting—criteria. These examples show that AHP is well suited for decision-making processes that require a balance between technical, operational, and economic factors, which supports its application in the analysis and selection of portable fundus cameras [29]. Thus, AHP stands out for its robustness, its ability to handle structured subjectivity, and its methodological clarity, which allows individual perceptions to be transformed into rational, measurable, and justifiable decisions [30].

III.2 PORTABLE FUNDUS IMAGING TECHNOLOGIES

Portable fundus cameras emerged as an alternative to traditional tabletop fundus imaging systems typically used in clinics and hospitals. These devices employ miniaturized optical systems combined with high-resolution digital sensors, enabling the capture of retinal images without pupil dilation or with minimal dilation, depending on the model [31] ,[32]. Among the leading globally recognized manufacturers are the Optomed Aurora, the Volk Pictor Plus, the Remidio Fundus on Phone, and compact models from Topcon, which—although more robust—offer versions with increased portability. These devices show significant differences in image quality, ergonomics, battery life, and integration with telemedicine software. Despite their advantages, challenges remain, such as balancing optical quality and portability, the high cost of some models, and the need for integration with electronic health record systems. Current trends point toward devices with enhanced connectivity, integration with artificial intelligence algorithms for automated pathology detection, and greater operational autonomy [33].

IV. METHODOLOGY

This study is characterized as applied research, quantitative and descriptive in nature, using the multicriteria AHP method for selecting portable fundus cameras. The methodological application followed a structured sequence beginning with a clear definition of the problem: selecting the most suitable device among the models available on the market. Next, the evaluation criteria were defined based on the literature and expert analysis: image quality, portability, battery life, cost, technical support, and compatibility with electronic record systems. The problem was then organized into a hierarchical structure composed of the objective at the top, the criteria at the intermediate level, and the evaluated alternatives at the base. Three representative market alternatives were considered: a model with high optical precision (A), an intermediate model with good cost-benefit ratio (B), and an ultracompact smartphone-based model (C). For each criterion, experts performed pairwise comparisons using Saaty's scale. The resulting matrices enabled the calculation of relative weights for the criteria and the priorities of the alternatives. The consistency ratio was calculated for all matrices to verify the reliability of the judgments. The analysis was carried out using automated spreadsheets developed in Excel, allowing extraction of eigenvectors, consistency indices, and the final ranking of alternatives.

IV.1 DATA COLLECTION AND PROCESSING

Data collection was conducted exclusively through documentary research and analysis of technical information provided by the manufacturers of the selected devices. This procedure is widely used in studies applying the AHP method, as it allows the gathering of objective data on characteristics, specifications, and functionalities of the evaluated alternatives, serving as the basis for building the comparisons. Pairwise comparisons between criteria and between alternatives were defined based on the analysis of the collected technical information, using Saaty's Fundamental Scale (1–9). The assigned values were organized in electronic spreadsheets (Microsoft Excel®), which were also used for calculating eigenvectors, normalization processes, verification of consistency indices, and obtaining the model's final priorities.

The data processing followed these steps:

• Construction of the judgment matrices
• Normalization of the matrices
• Calculation of priority vectors
• Calculation of the largest eigenvalue (λmax)
• Calculation of Consistency Indices

$$CI = \frac{\lambda \max - n}{n-1} \quad (1)$$

• **Calculation of the Consistency Ratio (CR):**

$$CR = \frac{CI}{RI} \quad (2)$$

Where RI is the Random Index presented by Saaty (1980). Only matrices with CR ≤ 0.10 were considered consistent, according to the classical recommendation in the literature [6],[20].

IV.2 DEFINITION OF CRITERIA

The definition of criteria considered three sources:

(i) specialized literature on teleophthalmology and diagnostic technologies;
(ii) international guidelines on ophthalmic image quality;
(iii) technical opinion of the interviewed experts.

The literature highlights that criteria such as image quality, technical support, portability, and digital compatibility are decisive in the adoption of portable diagnostic equipment [34]. Accordingly, six central criteria were defined for the analysis:

• **C1 – Image quality:** directly associated with diagnostic accuracy.
• **C2 – Portability:** fundamental for remote screenings.
• **C3 – Battery life:** relevant in fieldwork and regions without electrical infrastructure.
• **C4 – Cost:** including acquisition and cost-effectiveness.
• **C5 – Technical support and maintenance:** critical for operational continuity.
• **C6 – Digital compatibility:** integration with EHR, PACS, and telemedicine systems.

The choice of these criteria aligns with recommendations from technology assessment studies and the guidelines of the American Telemedicine Association.

IV.3 DETERMINATION OF WEIGHTS

To determine the weights of the criteria, the pairwise comparison method was applied according to the Fundamental Scale [35]. The study compared each criterion with the others in terms of relative importance, resulting in the matrix presented in Table 3. The weights (priority vector) were obtained through:

1. **Normalization of the AHP matrix:**

$$a'_{ij} = \frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}} \quad (3)$$

1. **Calculation of the weight vector (w_i):**

$$w_i = \frac{1}{n}\sum_{j=1}^{n} a'_{ij} \quad (4)$$

2. **Mathematical validation using the principal eigenvector:**

$$A\mathrm{w} = \lambda_{\max}\mathrm{w} \quad (5)$$

The final weights, presented in Table 4, highlighted the predominance of **C1 (0.30)** and **C5 (0.20)**, a finding consistent with previous studies on technology decision-making in clinical settings. The consistency of the judgments was verified by calculating the Consistency Ratio (CR), which in this study resulted in **CR = 0**, indicating full coherence of the comparisons.

IV.4 APPLICATION OF THE AHP METHOD

The application of the Analytic Hierarchy Process (AHP) in this study followed a structured set of steps that allowed the transformation of experts' subjective judgments into quantitative measures used to select the portable fundus camera most suitable for the clinical context. The process began with defining the alternatives, constructing the judgment matrix for the criteria, calculating the relative weights, assessing consistency, and, finally, synthesizing the overall priorities of the evaluated devices.

Initially, three commercially representative models of portable fundus cameras were selected based on their technical relevance and widespread use in teleophthalmology services. These models are presented in Table 2.

Table 2: Portable fundus camera models analyzed.

Commercial Model	Manufacturer	Main Features
Optomed Aurora	Optomed	High resolution, advanced ergonomics, suitable for teleophthalmology.
Volk Pictor Plus	Volk	Good image quality, widely used in screening.
Remidio Fundus on Phone	Remidio	Smartphone-attached system, extreme portability, and high digital integration.

Source: Authors, (2026).

Next, the six evaluation criteria were defined, representing the most relevant factors for choosing the device: image quality, portability, battery life, cost, technical support, and integration with electronic systems. Each criterion was described based on the literature and the evaluators' experience, as shown in Table 3.

Table 3: Evaluation Criteria.

Code	Criterion	Description
C1	Image Quality	Sharpness, resolution, and diagnostic capability of the fundus.
C2	Portability	Weight, size, and ease of transport.
C3	Battery Life	Continuous operating time without recharging.
C4	Cost	Purchase price and cost-effectiveness.
C5	Technical Support and Maintenance	Access to maintenance, training, and manufacturer assistance.
C6	Compatibility with Electronic Systems	Ability to integrate with EHR, PACS, and telemedicine platforms.

Source: Authors, (2026).

With the criteria defined, the pairwise comparison matrix was constructed using Saaty's Fundamental Scale (1 to 9). The experts assessed the relative importance of each pair of criteria, resulting in the matrix presented in Table 4.

Table 4: Pairwise Comparison Matrix of the Criteria.

	C1	C2	C3	C4	C5	C6
C1	1,00	2,00	3,00	2,00	1,50	3,00
C2	0,50	1,00	1,50	1,00	0,75	1,50
C3	0,33	0,67	1,00	0,67	0,50	1,00
C4	0,50	1,00	1,50	1,00	0,75	1,50
C5	0,67	1,33	2,00	1,33	1,00	2,00
C6	0,33	0,67	1,00	0,67	0,50	1,00

Source: Authors, (2026).

From this matrix, the priority vector (normalized principal eigenvector) was calculated, determining the weights of the criteria presented in Table 5.

Table 5: Relative Weights of the Criteria.

Criterion	Description	Relative Weight (wi)
C1	Image Quality	0.30
C2	Portability	0.15
C3	Battery Life	0.10
C4	Cost	0.15
C5	Technical Support and Maintenance	0.20
C6	Digital Compatibility	0.10
Sum		1.00

Source: Authors, (2026).

IV.5 ANALYSIS OF FUNDUS CAMERAS

The analysis of portable fundus cameras considered the three selected models (Table 1), evaluated across the six established criteria. The selection of these models was based on commercial availability, relevance in teleophthalmology programs, and established use in clinics and research, following recommendations from international studies. The models exhibited distinct characteristics, allowing for a robust multicriteria evaluation. The Optomed Aurora stood out for its high image quality and exemplary technical support; the Volk Pictor Plus demonstrated a balance between portability and cost; while the Remidio Fundus on Phone excelled in digital integration and extreme portability. Performance evaluation used scores from 1 to 9 (Saaty Scale), which were subsequently normalized and aggregated using the criteria weights. This procedure follows the additive synthesis process:

$$\text{"Overall Score of the Alternative"}\ A_k = \sum_{i=1}^{n} \quad (w_i \cdot p_{ik}) \tag{6}$$

V. RESULTS AND DISCUSSION

The results obtained through the application of the AHP method provided a clear understanding of the relative importance of each criterion in the selection of a portable fundus camera and the influence of these weights on the final ranking of the evaluated alternatives. Analysis of the pairwise comparison matrices revealed that the image quality criterion (C1) emerged as the most relevant in the decision-making process, corresponding to the highest weight among the analyzed criteria. This outcome is consistent with clinical practice, where diagnostic accuracy depends directly on the resolution and sharpness of the captured images. Next, the technical support and maintenance criterion (C5) stood out as an essential element to ensure the operational continuity of the equipment, reduce downtime, and provide qualified user assistance. The cost (C4) and portability (C2) criteria presented intermediate weights, reflecting their relevance, although they do not surpass the importance attributed to diagnostic quality and specialized support. The battery life (C3) and compatibility with electronic health record systems (C6) criteria had a lower relative impact on the decision, suggesting that, while relevant, they are perceived as secondary influencing factors compared to the technical aspects central to ophthalmic practice. These results reinforce patterns described in similar studies, where technological and support criteria carry greater weight in the acquisition of medical equipment. After defining the criteria weights, the alternatives were evaluated. The experts assigned scores from 1 to 9 to each model for each criterion, as presented in Table 6.

Table 6: Model Scores by Criterion (1–9 Scale).

Model / Criterion	C1	C2	C3	C4	C5	C6
A – Optomed Aurora	9	7	8	4	9	8
B – Volk Pictor Plus	8	8	7	5	8	7
C – Remidio Fundus on Phone	7	9	6	7	7	9

Source: Authors, (2026).

These scores were normalized and combined with the relative weights of the criteria, resulting in the overall priorities presented in Table 7.

Table 7: Overall Priorities of the Alternatives.

Model	**Description**	Overall Priority
A	Optomed Aurora	0,340
B	Remidio Fundus on Phone	0,334
C	Volk Pictor Plus	0,326

Source: Authors, (2026).

The analysis of the overall priorities highlights a significant balance among the evaluated models, indicating that all exhibit competitive performance. However, the Optomed Aurora (Model A) stood out as the most suitable alternative, achieving an overall priority of 0.340. This superior performance is primarily due to the combination of excellent image quality (C1) and robust technical support (C5), which are precisely the most influential criteria in the AHP model. Model C (Remidio FOP) achieved a very close score (0.334), benefiting mainly from its high portability and excellent digital compatibility. Model B (Volk Pictor Plus) obtained an overall priority of 0.326, favored for its good balance between image quality, portability, and cost, although it did not excel in the criteria deemed most critical by the experts. The consistency of the matrices was verified to ensure the reliability of the judgments. Considering the largest calculated eigenvalue ($\lambda max = 6.0$), the following was obtained:

$$CI = 0 \tag{7}$$

$$RI\ (n = 6) = 1{,}24 \tag{8}$$

$$CR = 0/1{,}24 = 0 \tag{9}$$

The CR value of 0 indicates perfect consistency in the pairwise comparisons, fully meeting the parameters defined by Saaty. This ensures that the experts' judgments were coherent and that the developed model is mathematically stable and reliable. The sensitivity analysis showed that small variations in the criteria weights did not significantly change the final ranking of the alternatives. This result highlights the robustness of the applied AHP model, reinforcing the reliability of the selection made. It is observed that technical and operational criteria, particularly those related to product quality and after-sales support, tend to exert a decisive influence on the selection of medical technologies. The integration of expert judgment, mathematical modeling, and structured analysis allowed the conclusion that the Optomed Aurora is the most suitable alternative for teleophthalmology services and clinical practice, offering the best balance between technical performance, specialized support, and operational applicability.

VI. CONCLUSION

This study aimed to apply the AHP method to support the selection of a portable fundus camera in the context of clinical practice and teleophthalmology, simultaneously considering multiple technical, operational, and economic criteria. Based on the definition of six main criteria—image quality, portability, battery life, cost, technical support, and digital compatibility—and the evaluation of three representative commercial models, it was possible to structure the problem into a clear hierarchy, develop pairwise comparison matrices, calculate relative weights, and synthesize the overall priorities of the alternatives. The results showed that image quality (C1) and technical support and maintenance (C5) were the most influential criteria in the decision-making process, followed by cost (C4) and portability

(C2), while battery life (C3) and digital compatibility (C6) had a relatively lower impact. This hierarchy is consistent with ophthalmic practice, where diagnostic reliability and the operational continuity of equipment are central factors for providing safe and effective care. Prioritization of technical and support criteria also aligns with previous studies on medical equipment selection using AHP or MCDA. In the synthesis of overall priorities, all three analyzed models demonstrated competitive performance. However, the Optomed Aurora (Model A) was identified as the most suitable alternative, with an overall priority of 0.340, closely followed by the Remidio Fundus on Phone (Model C) and the Volk Pictor Plus (Model B). This result indicates that, although the equipment is technically competitive, the combination of excellent image quality and robust technical support distinguishes Model A in the considered use context. The closeness of the scores also suggests that local decisions—such as budget constraints, regional availability of technical support, or institutional telemedicine strategies—may justify the selection of other models in specific scenarios, highlighting the usefulness of AHP as a tool adaptable to the priorities of each service. Consistency verification of the judgments, with a consistency ratio (CR) equal to zero for the criteria matrix, ensured the internal coherence of the comparisons and the mathematical stability of the adopted model. Sensitivity analysis indicated that moderate variations in criteria weights did not alter the ranking of alternatives, suggesting robustness of the solution obtained. These findings support the literature that identifies AHP as a transparent, auditable, and suitable method for supporting health technology acquisition decisions, particularly when multiple stakeholders and criteria must be considered in an integrated manner. The main contributions of this study are: (i) proposing and applying a specific AHP model for the selection of portable fundus cameras, adapted to the demands of teleophthalmology services; (ii) empirically demonstrating that image quality and technical support tend to prevail over strictly economic criteria in the decision-making process; and (iii) providing a replicable methodological framework that can be adjusted to include other models, additional criteria, or different expert profiles. From a practical perspective, the model can be incorporated into clinical engineering processes, procurement committees, or telemedicine programs to support purchase decisions in a systematic and justifiable way. However, some limitations should be acknowledged. The number of experts involved was limited, which may restrict the generalizability of the judgments, although internal consistency was satisfactory. In addition, the study considered only three commercial models, not covering the full spectrum of portable fundus cameras available on the international market. Important criteria such as life-cycle costs, real-world performance indicators (e.g., non-diagnostic exam rates, operator learning curve), and user experience metrics (usability, ergonomics) were not explicitly included in the model. For future research, it is recommended to expand the number of experts and devices evaluated, including fundus cameras from different price ranges and technological generations. Incorporating hybrid approaches, such as fuzzy AHP or integration with methods like TOPSIS, VIKOR, or PROMETHEE, may allow better handling of uncertainties and more complex preference scenarios. It is also suggested to include criteria related to integration with artificial intelligence algorithms for automated retinopathy detection and the evaluation of field clinical indicators, in order to bring the model closer to the real needs of remote ophthalmology screening and diagnostic programs. In this way, the present study constitutes an initial step in developing a systematic approach for selecting portable fundus cameras, paving the way for more comprehensive and integrated decision-support models in ophthalmic technologies.

VII. REFERENCES

[1] Alves, Daniela Osório. Desempenho de retinografia portátil com inteligência artificial no rastreio de retinopatia diabética na atenção primária em saúde. 2024.

[2] Faria, Thiago Pereira. Retinógrafo portátil no rastreamento de retinopatia diabética. Telemedicina e inteligência artificial. 2023.

[3] Kumar, Vijay; Paul, Kolin. Fundus imaging-based healthcare: Present and future. ACM Transactions on Computing for Healthcare, v. 4, n. 3, p. 1-34, 2023.

[4] Ahn, Seong Joon; Kim, Young Hwan. Clinical applications and future directions of smartphone fundus imaging. Diagnostics, v. 14, n. 13, p. 1395, 2024.

[5] Digkoglou, P.; Papathanasiou, J. Application of multiple criteria decision aiding in environmental policy-making processes. International Journal of Environmental Science and Technology, v. 22, n. 8, p. 6967-6982, 2025.

[6] Saaty, Thomas L. Analytic hierarchy process. In: Encyclopedia of operations research and management science. Springer, Boston, MA, 2013. p. 52-64.

[7] Campolina, Alessandro Gonçalves et al. Análise de decisão multicritério para alocação de recursos e avaliação de tecnologias em saúde: tão longe e tão perto? Cadernos de Saúde Pública, v. 33, p. e00045517, 2017.

[8] Alves, Nathalia Rocha. Substituição ou compra de equipamentos médico hospitalares pela aplicação do método AHP. 2020. Dissertação de Mestrado. Universidade Federal do Rio Grande do Norte.

[9] Bonatti, Rogerio Amaral. A gestão da informação e o processo decisório no setor energético: aplicação do método AHP na mensuração de critérios e alternativas. 2015.

[10] Rivas, René Ernesto García. Uso do método multicritério para tomada de decisão operacional tendo em conta riscos operacionais, à segurança, ambientais e à qualidade. Salvador: Dissertação de mestrado–UFBA/Programa de Engenharia Industrial, 2016.

[11] Castro, Wilkson Ricardo Silva. Modelo multicritério de apoio a decisão para o problema de contratação de fornecedores de equipamentos médicos: uma aplicação no Hospital Universitário Onofre Lopes (HUOL). 2020. Dissertação de Mestrado. Universidade Federal do Rio Grande do Norte.

[12] Silva, Lara Karine Dias. Proposta de sistemática para priorização de indicadores de afastamento em ambientes hospitalares na gestão de saúde e segurança do trabalho. 2023.

[13] Maia, Alexandre Peixoto et al. Ferramenta multicritério para avaliação da obsolescência: melhorando a gestão dos equipamentos médico-assistenciais de um hospital universitário de Minas Gerais. 2021.

[14] Yang, Chai et al. Improving hospital based medical procurement decisions with health technology assessment and multi-criteria decision analysis. INQUIRY: The Journal of Health Care Organization, Provision, and Financing, v. 58, p. 00469580211022911, 2021.

[15] Ferreira, Ana Rita da Costa. Seleção e avaliação de fornecedores: estudo de caso de uma empresa do setor dos dispositivos médicos. 2020. Dissertação de Mestrado. Universidade do Minho (Portugal).

[16] Medeiros, Cátia. O uso do recurso de diagnóstico por imagem como indicador de processo e eficiência do sistema de saúde. 2019.

[17] Favarato, Arthur Pinheiro. Comparação da identificação e referenciamento da retinopatia diabética por retinografia portátil baseado em smartphone e exame de biomicroscopia de fundo. 2023. Tese de Doutorado. [sn].

[18] Checon, Maria Eduarda Smarzaro et al. Inteligência artificial na oftalmologia: Diagnóstico preciso, acesso ampliado e novos desafios. Journal of Medical and Biosciences Research, v. 2, n. 4, p. 771-779, 2025.

[19] Ahn, Seong Joon; Kim, Young Hwan. Clinical applications and future directions of smartphone fundus imaging. Diagnostics, v. 14, n. 13, p. 1395, 2024.

[20] Saaty, Thomas L. Analytic hierarchy process. In: Encyclopedia of operations research and management science. Springer, Boston, MA, 2013. p. 52-64.

[21] Schmidt, Angela Maria Atherino et al. Processo de apoio à tomada de decisão abordagens: AHP e Macbeth. 1995.

[22] Saaty, Thomas L. Basic theory of the analytic hierarchy process: How to make a decision. Revista de la Real Academia de Ciencias Exactas Fisicas y Naturales, v. 93, n. 4, p. 395-423, 1999.

[23] Campos, Ana Jarvirs de Melo. Análise de critérios de decisão em estratégia de arranjos produtivos locais: um estudo aplicando o processo de hierarquia analítica (AHP) no setor de cerâmica. 2009.

[24] Rodrigues, Claudiane Fonseca. Análise comparativa entre os métodos decomposição em valores singulares e análise de componentes principais envolvendo matrizes esparsas de grande porte. 2011.

[25] De Avila, Anderson Roberto Sevage. Um Estudo Sobre os Conceitos Matemáticos que Sustentam o Método de Análise Hierárquica (AHP). 2023.

[26] Nepomuceno, Raul Carneiro. O controle da subjetividade nas decisões judiciais em casos de colisão entre direitos fundamentais: hermenêutica, método e a tensão entre o racional e o irracional. 2019.

[27] Ribeiro, Maitê Garcia Cruz. Análise da sustentabilidade no setor de distribuição de energia elétrica no Brasil. Rio de Janeiro, Brasil, 2011.

[28] Santos, Gideão Oliveira dos. Proposta de métrica para avaliação de desempenho profissional. 2023.

[29] Araujo, Isabela Carolina Bertolo de. Aplicação do método AHP em uma empresa de desenvolvimento de softwares para a seleção de prestadores de serviço. 2024.

[30] Georg, Marcus Aurélio Carvalho. Proposta de modelo de mensuração de apetite a riscos cibernéticos: uso do método AHP e da estrutura básica de segurança cibernética.

[31] Titoneli, Carolina Carneiro. Validação clínica de um retinógrafo portátil conectado a um smartphone na avaliação da cabeça do nervo óptico. 2021. Tese de Doutorado. Universidade de São Paulo.

[32] Faria, Thiago Pereira. Retinógrafo portátil no rastreamento de retinopatia diabética. Telemedicina e inteligência artificial. 2023.

[33] Sartori, Landa Carretero Nunes Marques. A tecnologia usada como vantagem contra a criminalidade: uma análise da evolução tecnológica das forças de segurança, com um recorte para a Guarda Municipal de Vila Velha. 2024. Tese de Doutorado. brasil.

[34] Coelho Filho, Manuel Pereira. Análise econômica referente a adoção de ferramentas de inteligência artificial no auxílio à tomada de decisão clínica aplicadas à fase de diagnóstico. 2024. Tese de Doutorado. Universidade de São Paulo.

[35] Marchezetti, Ana Lúcia; Kaviski, Eloy; Braga, Maria Cristina Borga. Aplicação do método AHP para a hierarquização das alternativas de tratamento de resíduos sólidos domiciliares. Ambiente Construído, v. 11, p. 173-187, 2011.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 465-482. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3098

RESEARCH ARTICLE OPEN ACCESS

A SYSTEM FOR DETECTING NON-CONFORMITIES IN INDUSTRIAL MANUFACTURING PROCESSES IN THE AUTOMOTIVE SECTOR USING FUZZY LOGIC

Roberto Ferreira de Lima[1]*, Helder Kiyoshi Miyagawa[2] and Rivanildo Duarte Almeida[3]

[1]Federal Institute of Amazonas – IFAM. Manaus – Amazonas, Brazil

[2]The Institute of Technology of the Federal University of Pará (PPGEP-ITEC-UFPA). Avenida Augusto Correa, 01. Belém, Pará – Brazil.

[3]Instituto de Tecnologia e Educação Galileo da Amazônia. Manaus – ITEGAM. Manaus, Amazonas - Brasil

[1]https://orcid.org/0009-0000-3771-8807, [2]http://orcid.org/ 0000-0001-9346-4696, [3]http://orcid.org/ 0000-0001-9454-2393

Email: *roberto.lima@ifam.edu.br, helderkm@ufpa.br, duartiot@gmail.com

ARTICLE INFO

Article History

Received: December 10, 2025

Revised: December 20, 2025

Accepted: January 1, 2026

Published: January 31, 2026

Keywords:

Fuzzy Logic,

Non-Conformities,

PPM,

Process Control, Industry.

ABSTRACT

Increasing industrial competitiveness demands stable and continuously monitored production processes. This work presents a Non-Conformance Detection System based on Fuzzy Logic, applied to quality control in manufacturing. Using the PPM (Parts Per Million) indicator, the system evaluates process performance and identifies the risk level of non-conformities. Scenarios with PPM between 20 and 11,000 were analyzed, covering conditions from ideal to critical situations. The fuzzy model, composed of triangular and trapezoidal membership functions, expert rules, and centroid defuzzification, showed smooth transitions and greater sensitivity compared to deterministic methods. The results confirm its practical applicability in the automotive industry, assisting in risk identification, rapid decision-making, and continuous process improvement.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Quality control in industrial manufacturing systems has been a recurring theme since the advent of mass production in the Industrial Revolution, when the standardization of processes and products became essential to meet the growing market demand. In recent decades, with globalization, Industry 4.0, and the increased complexity of production processes, fault detection and non-conformity prevention have become critical factors for the competitiveness and sustainability of organizations [1]. The problem under study is directly related to the non-detection of non-conformities in industrial manufacturing processes, which can lead to rework, waste of resources, increased costs, and, above all, compromised customer satisfaction. The literature points out that the absence of effective monitoring mechanisms can lead to the repetition of errors and the loss of reliability of the production system [2]. The current demands of Industry 4.0 require increasingly precise, flexible, and sustainable processes. The absence of effective fault detection mechanisms compromises productivity and quality, negatively impacting the competitiveness of companies. In this sense, the adoption of techniques based on Artificial Intelligence and Fuzzy Logic is highly aligned with the needs of the sector, as it allows for the handling of uncertainties and variabilities common in industrial processes. In this context, several methodologies and theories have been used to minimize such problems. Among the most relevant approaches are statistical process control techniques, Lean Six Sigma methodologies, and, more recently, intelligent systems based on computational intelligence [3], [4]. Among these techniques, fuzzy logic stands out for its ability to handle imprecise and uncertain information, offering a model closer to human decision-making. The objective of this dissertation is to develop a system for detecting nonconformities in industrial manufacturing processes, based on fuzzy logic to integrate process signals and quality criteria, reducing the non-detection rate and the response time of inspections.

II. INDUSTRIAL MANUFACTURING SYSTEM

Production systems are defined as socio-technical systems responsible for transforming inputs into products through structured and organized processes [5]. Production involves interrelated technical and human functions, based on systems theory and cybernetics, considering human, social, technical, and organizational subsystems integrated into the environment [6]. Value creation requires organized processes and operations such as machining, assembly, testing, and transportation, performed by people and equipment [7]. With the development of the Toyota Production System, a holistic approach to Lean Production Systems (LPS) emerged, combining strategic principles, operational methods, and tools aimed at eliminating waste and continuous improvement [5]. These systems guide all processes towards the customer and seek the systematic reduction of activities that do not add value [8]. Manufacturing has evolved through four industrial revolutions: mechanization; adoption of new energy sources and combustion engines; automation with electronics and computing; and, more recently, Industry 4.0, marked by cyber-physical systems, IoT, cloud computing, and cognitive resources [9]. Companies aligned with Industry 4.0 are considered "smart," with flexible, precise, digitized, and logistically integrated systems[10]. Digital transformation occurs in four stages—digitization, virtualization, connectivity, and autonomization—leading to new scenarios of human-machine cooperation in smart factories [5].

II.1 INTELLIGENT MANUFACTURING SYSTEM

Intelligent manufacturing is an advanced production model that uses information technologies, sensors, data analysis, and intelligent systems to optimize all stages of a product's life cycle, from design to final integration. Its goal is to increase the efficiency, quality, and competitiveness of companies in dynamic markets [11]. The advancement of automation, driven by the evolution of machines and computers, has led to the emergence of concepts such as flexible cells, computer-integrated manufacturing, and, finally, intelligent manufacturing. Since the 1990s, international programs—especially in Japan, the USA, and Europe—have driven research and cooperation to develop this new productive paradigma [12]. Smart manufacturing integrates communication technologies, data science, systems engineering, and automation to connect people, machines, and processes. This integration occurs through detection, interconnection, analysis, learning, and autonomous decision-making, allowing the optimization of three central elements (people, management, and technology) and five fundamental flows (information, logistics, capital, knowledge, and service) [13]. According to [14], smart manufacturing relies heavily on data, sensors, and advanced analytics, integrating Operational Technology (OT) and Information Technology (IT). According to [12] organizes the concept into six fundamental pillars:

1) Manufacturing technologies and processes – including additive manufacturing, new materials, hybrid processes, and smarter robots.
2) Materials – use of new materials, biomaterials, and resource recovery at the end of the life cycle.
3) Data – massive collection and advanced analysis, basis for predictive models and knowledge preservation.
4) Predictive engineering – creation of digital models to predict future behaviors and support strategic decisions.
5) Sustainability – development of sustainable products and processes, valuing remanufacturing, reuse, and impact reduction.
6) Resource sharing and networking – cooperation between companies to share equipment, software, and knowledge.

Artificial Intelligence is essential in SMI, enabling real-time monitoring, automated decision-making, intelligent scheduling, and increasingly advanced human-machine integration [11]. In Industry 4.0, SMI uses service-oriented architecture (SOA) and platforms connected via the Internet, forming an integrated ecosystem where technical, managerial, and organizational processes operate collaboratively, flexibly, and reconfigurably.

II.1.1 Sistema de Manufatura Flexivél

A flexible manufacturing system (FMS) can be defined as a computer-controlled production system capable of processing a variety of part types, characterized by the ability of production to be adapted to the production of multiple parts, small or medium batches, and there are several flexible manufacturing systems categorized according to the target process being improved. These categories are addressed below [9]:

a) Variant flexibility: Ability to manufacture or assemble more variants of a product. (Product flexibility)
b) Quantity flexibility: Ability to adapt the production system to fluctuating volumes.
c) Technology flexibility: The ability of the manufacturing and assembly system to be used for a range of technologies.
d) Successor flexibility: Ability to use equipment or parts also for future products.

II.2 QUALITY IN THE INDUSTRIAL MANUFACTURING SYSTEM

Quality in industrial production involves the ability of a production system to generate goods that meet or exceed technical standards, requirements, and customer expectations, from conception to delivery [15] and [16]. This concept encompasses conformity, reliability, and continuous improvement throughout the production process.

II.2.1 Quality Management System (QMS)

The Quality Management System (QMS) promotes the standardization of processes with a focus on continuous improvement and customer value [15]. Implementing a QMS means documenting procedures, monitoring performance, and correcting deviations systematically, ensuring that all products meet the established requirements.

II.2.2 Evolution towards Industry 4.0 and Quality 4.0

The transition to Industry 4.0 has transformed quality approaches: integrated systems and connectivity have become fundamental, characterizing what is now called Quality 4.0 [16]. Technologies such as automation, intelligent sensors, and real-time data analysis allow for greater efficiency and reliability [17]. Digital integration enables preventive actions and data-driven decisions, increasing the competitiveness of the industry.

II.2.3 Quality Tools and ISO 9001

The implementation of ISO 9001:2015 requires the establishment of an effective QMS, capable of documenting processes, reducing failures, and ensuring conformity [18]. In the Brazilian context, the application of tools such as PDCA, 5S, and statistical methods proves essential to reduce costs, rework, and improve customer satisfaction [19].

II.2.4 Quality Tools

Quality tools are essential instruments for monitoring, analyzing, and improving processes in industrial environments. Below are the main ones:

a) Pareto Diagram

The Pareto Diagram organizes defects in descending order of frequency, allowing the identification of the most relevant causes according to the Pareto principle, which indicates that a few factors are responsible for the majority of problems. Thus, prioritizing the main causes allows for a greater impact on process improvement [20]. Diagram is shown in the figure 1.

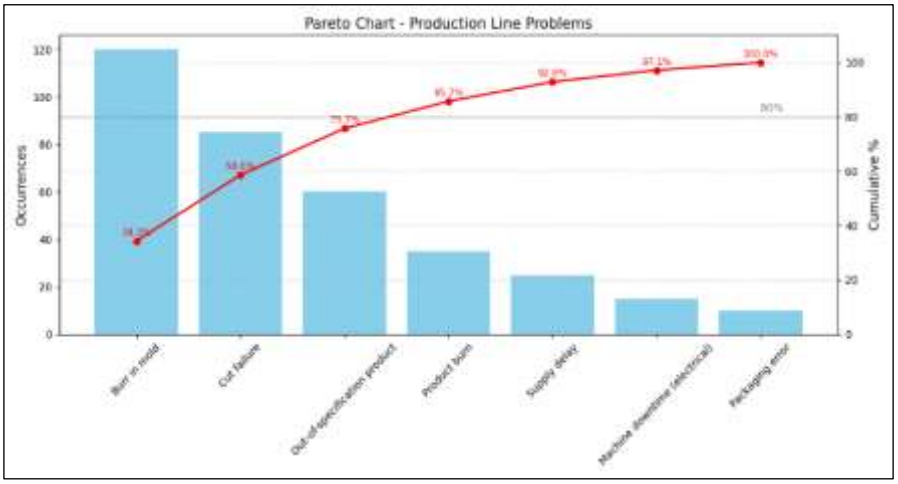


Figure 1: Pareto Diagram.
Source: Authors, (2026).

b) Control Chart

The Control Chart is used to monitor the stability of a process over time. It presents upper and lower control limits, as well as a centerline. When the points remain within these limits, the process is considered stable; values outside the limits indicate a need for intervention [21]. The control chart is shown in the figure 2.

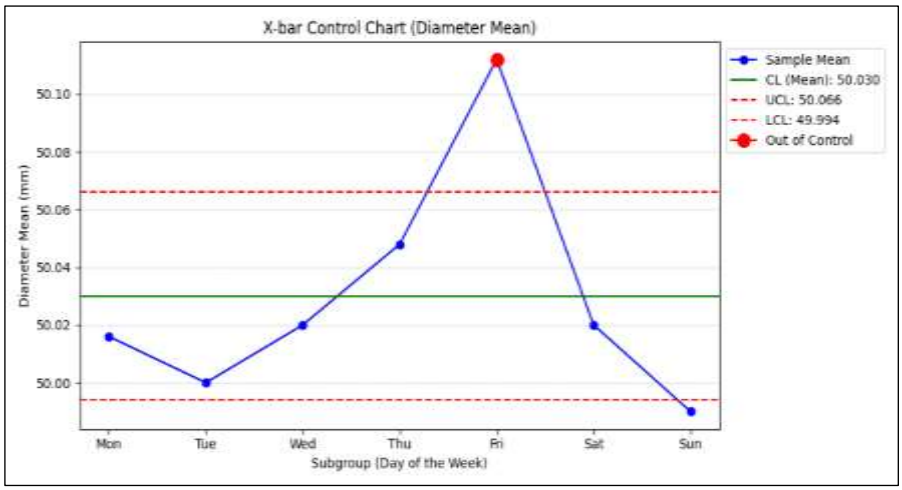


Figure 2: Iron bar production.
Source: Authors, (2026).

c) Ishikawa Diagram

The Ishikawa Diagram, or cause-and-effect diagram, assists in the systematic identification of the possible causes of a problem, organizing them into categories such as methods, machines, manpower, materials, measurements, and environment. This facilitates understanding the origins of a nonconformity [22]. The Ishikawa Diagram is shown in the figure 3.

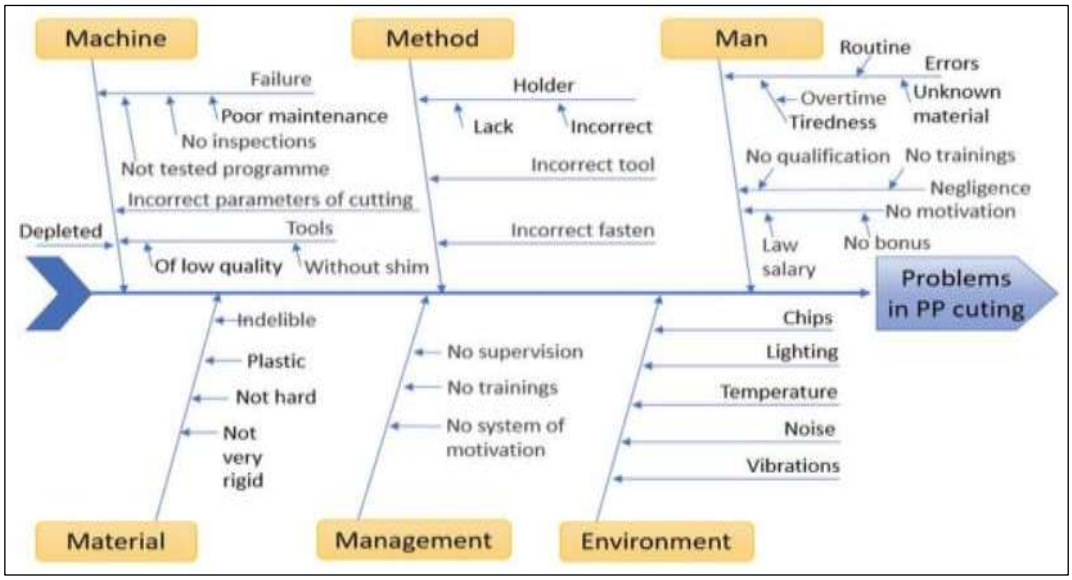


Figure 3: Ishikawa diagram for polypropylene cutting.
Fonte: [22].

d) Flowchart

Finally, the flowchart graphically represents the steps of a process, allowing you to visualize the logical sequence of activities, identify critical points, and support operational improvements. This tool is fundamental for standardizing procedures and understanding the production flow clearly. The figure below shows an example of a flowchart.

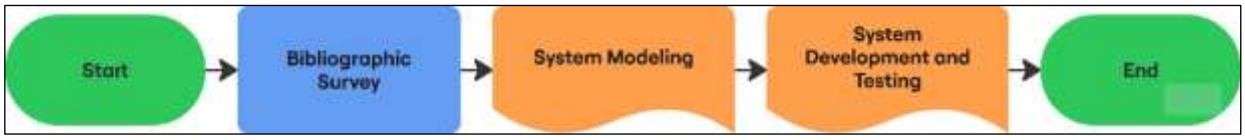


Figure 4: Methodological flow.
Source: Authors, (2026).

II.3 TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

This topic will cover the most commonly used techniques in industrial processes. Fuzzy Logic, Simple Linear Regression, Decision Trees, and Artificial Neural Networks are widely used in intelligent systems for decision support, process control, and modeling of industrial phenomena.

II.3.1 Fuzzy Logic

Fuzzy Logic allows working with imprecision and subjectivity, representing variables by degrees of membership between 0 and 1. Unlike traditional Boolean logic, it allows an element to partially belong to several sets. This approach is applied in different stages of manufacturing, such as temperature control, defect reduction, and injection parameter adjustment. The operation of a fuzzy system involves three fundamental steps: fuzzification, which converts numerical values into linguistic terms; inference, which applies "If-Then" rules defined by experts; and defuzzification, which transforms the fuzzy result into a precise numerical value. The most common membership functions are triangular, trapezoidal, and Gaussian, with the triangular function being the most used due to its simplicity and computational efficiency [23]. The linguistic variable is a variable whose elements are names of fuzzy sets. Its main function is to provide a way to characterize the complexity of phenomena and the lack of clarity. This allows the treatment of more complex systems to be analyzed using traditional mathematical terms [24]. The most common forms of fuzzy membership functions are: triangular fuzzy, trapezoidal fuzzy, and Gaussian fuzzy. See figures 5, 6, and 7, respectively.

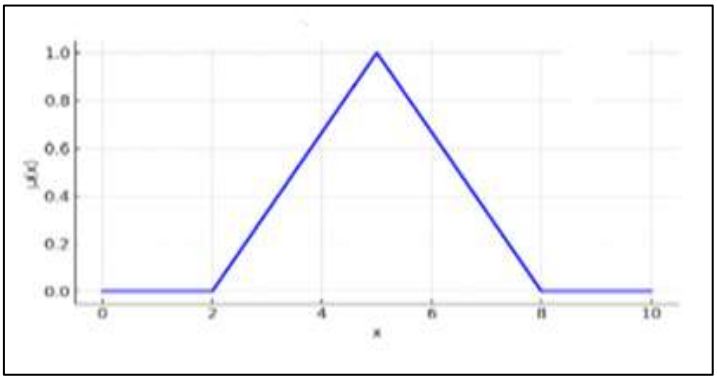


Figure 5: Fuzzy Triangular Membership Function.
Source: Authors, (2026).

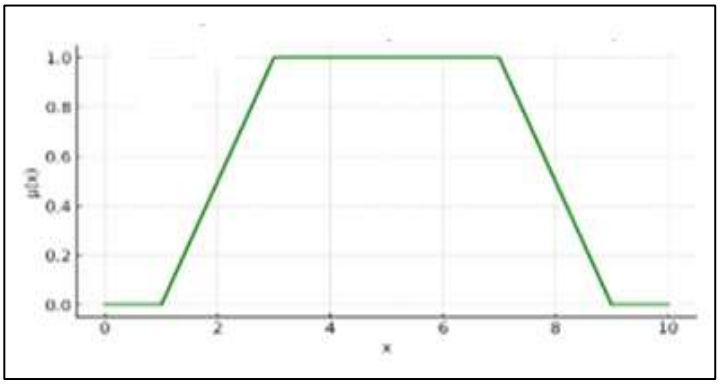


Figure 6: Fuzzy Trapezoidal Membership Function.
Source: Authors, (2026).

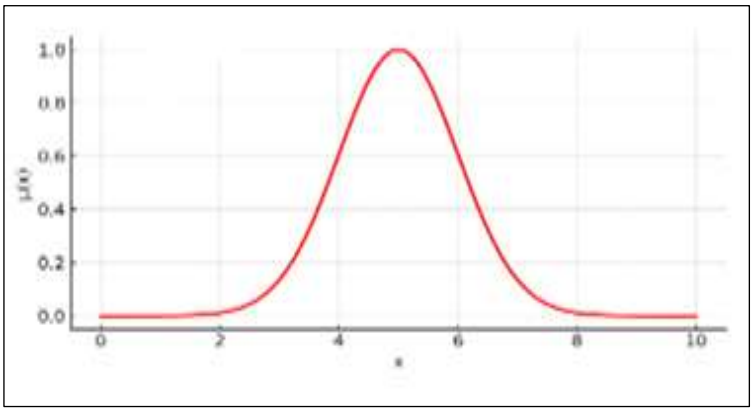


Figure 7: Fuzzy Gaussian Membership Function.
Source: Authors, (2026).

II.3.2 Simple Linear Regression

Simple linear regression models the relationship between two variables one explanatory and one response using a straight line defined by the intercept (α) and slope (β) parameters. This model allows for predicting future behaviors and identifying trends, being widely applied in production processes, demand forecasting, costs, or operational performance. The error term (ei) represents the random variation not explained by the equation. Despite its simplicity, this technique is effective when there is a linear correlation between the variables and provides a clear interpretation of how changes in x affect y. Figure 8 shows the graph of a linear function [25].

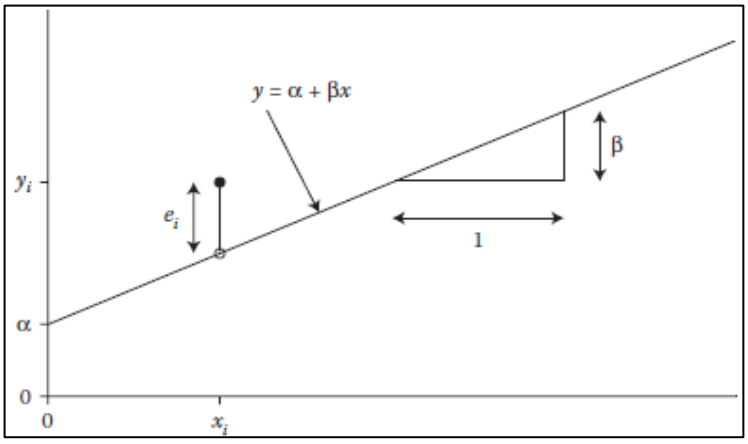


Figure 8: Simple linear regression model.
Source: [25].

II.3.3 Decision Tree

A Decision Tree is a supervised learning method used for classification or regression. It functions like a flowchart, where each node represents a decision test, each branch is a result, and each leaf indicates a class or prediction. Among its main advantages are ease of interpretation, low computational cost, and the ability to handle imperfect or redundant data. The construction of the tree involves the recursive partitioning of data until stopping criteria are met, followed by a pruning step, which avoids overfitting and improves the generalization of the model. This type of technique is widely used in medical applications, risk analysis, quality analysis, and industrial decision-making. Figure 9 depicts an example of a decision tree designed for classification, obtained by running the J48 (C4.5) algorithm applied to the Weka "contact lenses" dataset [26].

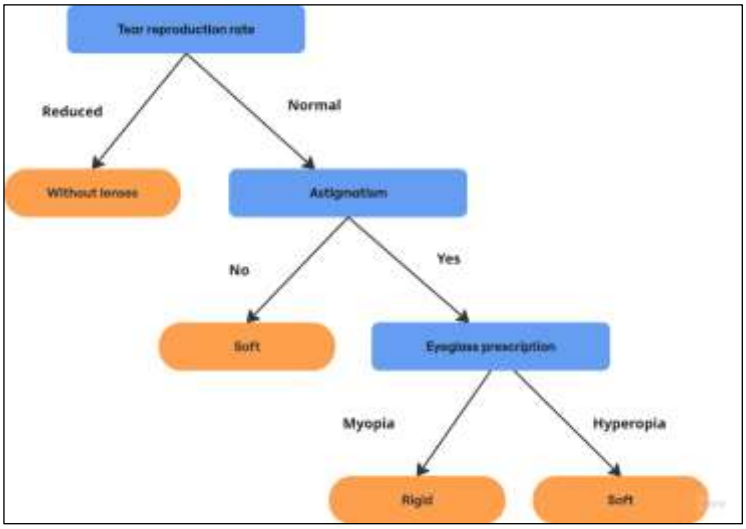


Figure 9: An example of a decision tree.
Source: Authors, (2026).

II.3.4 Artificial Neural Networks

Artificial Neural Networks are models inspired by the functioning of the human brain and are capable of learning complex patterns. They are formed by layers of interconnected artificial neurons that process information by multiplying inputs by weights and applying activation functions. The learning process occurs through backpropagation of error, adjusting the weights to improve accuracy. ANNs are especially effective in problems that require pattern recognition, time series prediction, complex classification, and decision-making based on large volumes of data. Due to their flexibility, they are widely used in computer vision, industrial automation, intelligent control, and predictive systems [27]. Figure 10 shows the neuron function.

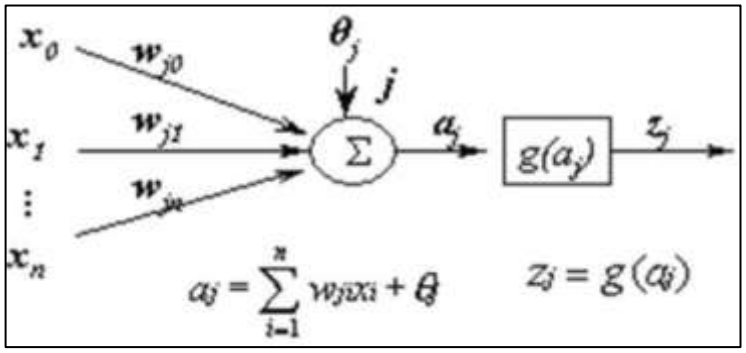


Figure 10: Neuron function.
Source: [27].

III. MATERIALS AND METHODS

For the development of the proposed system, the following computational tools and resources were used:

a) Programming language: Python;
b) Development environment: Visual Studio Code (VS Code);
c) Python libraries: NumPy, Matplotlib, and scikit-fuzzy (for fuzzy modeling and inference);
d) Auxiliary software: Miro (for creating flowcharts and modeling diagrams);
e) Simulation environment: Windows 11 operating system.

The use of the Python language is justified by its wide application in scientific research and the availability of libraries specialized in fuzzy logic and data analysis, which facilitates the implementation of intelligent systems and the handling of large volumes of information. The use of the Visual Studio Code (VS Code) IDE is justified because it is the most widely used IDE for software and embedded projects in the world. The methodology proposed in this dissertation is based on applied and experimental research, since its objective is the development and validation of a computational system aimed at detecting nonconformities in industrial manufacturing processes. The applied nature of the study is justified by seeking a solution to a real problem in the industrial area, which is the early identification of failures in the production process through monitoring the Parts Per Million (PPM) indicator. In addition, the research has an experimental and exploratory character, as it involves the conception, implementation, and evaluation of a computational prototype based on fuzzy logic, developed to simulate typical conditions of an industrial environment. The experimental approach allows testing the system's behavior in the face of different input scenarios, making it possible to analyze its inference and response capacity in nonconformity situations. The technique used in the methodology of this dissertation is the computational intelligence technique Fuzzy Logic.

Fuzzy logic is a theory proposed by Lofti Zadeh in 1965, which is based on sets that have degrees of membership and are encoded in unit intervals; that is, they have a membership function whose values are compatible with these sets in degrees of membership from 0 to 1, differing from classical logic theory, since the data are complex, that is, a piece of data has membership in a set, "all or nothing", which is based on sets whose inclusion is resolved in degrees of "true or false", 0 or 1, all or nothing; unlike fuzzy logic, where a piece of data is part of a set to a certain extent, to a higher degree (0), sometimes to a medium degree (0.5) or can increase or decrease its vagueness, and may even be empty [28].

III.1 RESEARCH STAGES

The development of this dissertation was structured in three main stages. These are:

a) Bibliographic survey

This stage consisted of conducting theoretical and documentary research on the main concepts related to industrial quality, process control, performance indicators (especially the Parts Per Million KPI), as well as the fundamentals of fuzzy logic and its applications in decision support systems. Academic sources, scientific articles, dissertations, and books addressing the topic were consulted in order to support the conceptual and technical development of the proposed system.

b) Fuzzy system modeling

In this phase, the input and output parameters of the fuzzy system were defined, based on variables relevant to the evaluation of the quality of the production process. The modeling involved the elaboration of membership functions, the definition of fuzzy rules (e.g., If cost LOW and benefit HIGH then cost-benefit HIGH), and the structuring of the inference mechanism, in order to approximately represent human reasoning in making decisions about the conformity or non-conformity of the process. The Miro software was used to design the flowcharts representing the system's operational logic and data flow, contributing to the visualization of the decision-making process.

c) System Development and Testing

After modeling, software was developed in Python, using the Visual Studio Code (VS Code) IDE, to implement the non-conformity detection and KPI monitoring system Parts Per Million (PPM). The program was structured to collect and process simulated production data, apply the fuzzy model, and return a monitoring graph of the indicator, allowing the user to identify failure trends and make corrective decisions more quickly. Computational simulation tests were performed, in which different data scenarios were entered into the system to evaluate its robustness, coherence of fuzzy inferences, and reliability of results. This step was essential to verify the adherence of the proposed model to the expected behavior in a real industrial environment.

Figure 11 shows the research steps represented in a methodological flow adopted to achieve the established objectives.

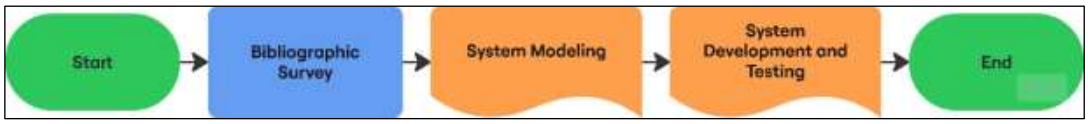


Figure 11: Methodological flow.
Source: Authors, (2026).

The system validation was performed in a simulated manner, using datasets representative of an industrial manufacturing process. The simulations allowed us to verify the consistency of the fuzzy system outputs in relation to the input conditions and to analyze the correlation between the KPI PPM (Parts Per Million) values and the compliance levels identified by the model. The results were evaluated based on performance criteria related to the accuracy in detecting non-conformities, the stability of the inferences, and the clarity of the information presented in the monitoring graph. This analysis made it possible to verify the viability of the system as a decision support tool in production processes.

III.2 FUZZY SYSTEM DEVELOPMENT

This topic presents all the steps used in the implementation of the fuzzy system based on Fuzzy Logic to monitor the Parts Per Million (PPM) indicator in manufacturing processes in the automotive supply chain. The objective is to transform numerical process data into an intelligent risk interpretation, capable of adaptively anticipating nonconformities. The introduction of this fuzzy system allows for the evaluation of quality levels even in imprecise scenarios or with natural process fluctuations, something that traditional methods cannot adequately represent.

III.2.1 General system architecture

The application was structured in a logical flow composed of four main steps:

a) Input: actual value of the PPM indicator collected in the industrial process.
b) Fuzzy processing: composed of fuzzification, rule base, and inference.
c) Defuzzification: transformation of the fuzzy result into a numerical index.
d) Output: Risk Index (0 to 100), indicating the degree of compliance of the process.

The system architecture can be represented as shown in figure 12.



Figure 12: System architecture flowchart.
Source: Authors, (2026).

III.2.2 Definition of the fuzzy input variable – PPM

The PPM indicator is widely used in industry to measure the number of defective parts in relation to the total produced. Table x abaixo presents the limits for PPM indicators according to the IATF 16949 standard.

Table 1: Limits for PPM indicators – IATF 16949 standard.

Level	PPM Range	Interpretation
Excellent	0 – 100 ppm	Very capable process
Good	100 – 500 ppm	Proper control
Acceptable	500 – 1.000 ppm	As expected, but with caution.
Warning	1.000 – 5.000 ppm	It requires investigations and corrective actions.
Bad	> 5.000	Process out of control.
Critic	>10.000	High risk – possible customer account suspension.

Source: Authors, (2026).

Based on this classification, six fuzzy sets were defined for the PPM:
Unit: parts per million (PPM). Universe considered: 0 20,000 PPM (sufficient to cover >10,000).

a) **Excellent:** 0 – 100 ppm (Very capable process)
b) **Good:** 100 – 500 ppm (Proper control)
c) **Acceptable:** 500 – 1.000 ppm (As expected, but with caution.)
d) **Warning:** 1.000 – 5.000 ppm (It requires investigations and corrective actions.)
e) **Bad** > 5.000 ppm (Process out of control)
f) **Critic:** > 10.000 ppm (High risk – possible customer account suspension)

Each set represents a region of the PPM space with smooth transitions between boundaries, a fundamental characteristic of fuzzy systems.

III.2.3 Membership functions of the PPM variable

The membership functions were defined to reflect industrial logic, using triangular and trapezoidal functions.

a) **Excellent:** — trapezoidal: (a=0, b=0, c=25, d=100) - High relevance in very low PPMs; decreases down to 100.
b) **Good:** — triangular: (100, 300, 500) - Peak at 300, base 100–500.
c) **Acceptable:**— triangular: (500, 750, 1000) - Peak at 750, base 500–1000.
d) **Warning:** — triangular: (1000, 3000, 5000) - Peak at, base 1000–5000.
e) **Bad** — triangular: (5000, 7500, 10000) - Peak at 7500, base 5000–100000.
f) **Critic:** — trapezoidal: (10000, 11000, 12000, 12000) - Starts to increase from 10k, becomes strong above 15k.

III.2.4 Fuzzy output variable – Risk index (0–100)

The output variable does not represent PPM values. It represents a fuzzy interpretation of the current process risk. For this reason, the output is defined on a universal scale of 0 to 100, widely used in risk analysis systems. Output universe: 0 …100 (0 = zero risk; 100 = maximum/critical risk). Fuzzy output sets (with triangular/trapezoidal shapes):

a) **Very Low** — tri (0, 10, 20)
b) **Low** — tri (10, 25, 40)
c) **Moderate** — tri (30, 50, 70)
d) **High** — tri (60, 75, 85)
e) **Very high** — tri (80, 90, 95)
f) **Critical** — trap (90, 95, 100, 100)

III.2.5 Construction of fuzzy rules

The rules were constructed based on a direct interpretation of the PPM classification table:

IF 0 ≤ PPM <= 100 It is **Excellent** → Risk Index Is **Very Low**
IF 100 ≤ PPM <= 500 It is **Good** → Risk Index Is **Low**
IF 500 ≤ PPM <= 1,000 It is **Acceptable** → Risk Index Is **Moderate**
IF 1,000 ≤ PPM <= 5,000 It is **Alert** → Risk Index Is **High**
IF PPM >= 5,000 It is **Bad** → Risk Index Is **Very High**
IF PPM >= 10,000 It is **Critical** → Risk Index Is **Critical**

These rules ensure that the system responds gradually, avoiding abrupt jumps in the ranking.

III.2.6 Fuzzy Inference Process

The Mamdani inference method was used, a method traditionally employed in industrial applications due to its robustness and interpretability.

AND operator used: minimum (typical in Mamdani).
Implication: clipping of the consequent's moiety by the degree of activation (minimum).
Output Aggregation: maximum (take union / sup) between the clipped sets.
Flow for a specific PPM:

a) Calculate the PPM membership level in each input set.
b) For each rule, obtain the activation level = membership of the antecedent.
c) Clip the MF of the corresponding output set at the activation level.
d) Aggregate (max) all clipped output MFs to obtain the aggregated fuzzy MF of the risk index.

III.2.7 Defuzzification

To transform the output fuzzy set into a numerical value, the following method was used:

Centroid – Center of Gravity (COG)
This method calculates the "center of mass" of the resulting area, providing a final risk index between 0 and 100.

IV. RESULTS AND DISCUSSIONS

This topic presents the results obtained from the implementation and testing of the fuzzy system developed for detecting nonconformities in industrial manufacturing lines. The results were organized, interpreted, and compared with industrial quality control practices, especially the use of metrics such as PPM (Parts Per Million), widely used for monitoring process performance. Obtaining real PPM data for the project was not possible due to the difficulty of accessing industrial information, which is usually confidential. To enable testing and maintain methodological consistency, simulated data were used, realistically constructed to represent different process situations. This approach allowed the system to be validated and its operation demonstrated, making the use of simulated data a necessary and justifiable alternative. The simulations were performed using PPM values obtained or defined in representative industry scenarios: 20, 150, 400, 800, 2500, 6000, and 11000. Each value was processed by the fuzzy system, resulting in an aggregate risk level and indicating the probability of the process exhibiting compliance, intermediate risk, or severe non-compliance.

The tests were conducted with the objective of verifying the fuzzy system's ability to:

a) Interpret PPM levels reported by the production process;
b) Add expert knowledge through fuzzy rules;
c) Automatically classify the degree of non-compliance risk;
d) Indicate recommended actions for each situation.

IV.1 IMPLEMENTAÇÃO COMPUTACIONAL DO SISTEMA FUZZY

This topic presents the implementation of a fuzzy system to monitor the PPM indicator in a production process in the automotive industry. The development was carried out in Python, chosen for its simplicity, readability, versatility, and large ecosystem of libraries, which facilitates prototyping and maintenance. The IDE used was Visual Studio Code, a lightweight, free, and highly extensible editor that offers features such as debugging, support for multiple languages, Git integration, and a built-in terminal. After defining all the implementation steps in topic 3.2.3, the execution of the system developed in Python is presented. Figure 13 exemplifies the behavior of the input variable in the system, demonstrating that the PPM variable has six membership functions (PFs), two trapezoidal (Excellent and Critical) and four triangular (Good, Acceptable, Alert, and Bad).

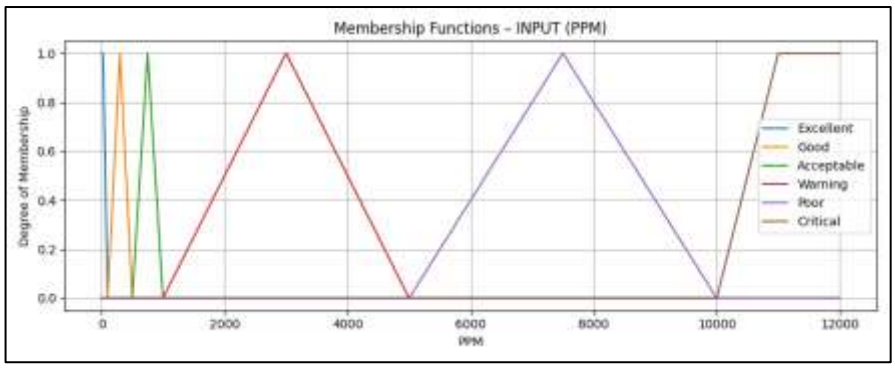


Figure 13: Input membership functions.
Source: Authors, (2026).

Figure 14 illustrates how the output variable is admitted in the proposed fuzzy system, considering that the Risk Index output variable, related to the PPM input, also has six membership functions (PFs), one trapezoidal (Critical) and five triangular (Very low, Low, Moderate, High, Very High).

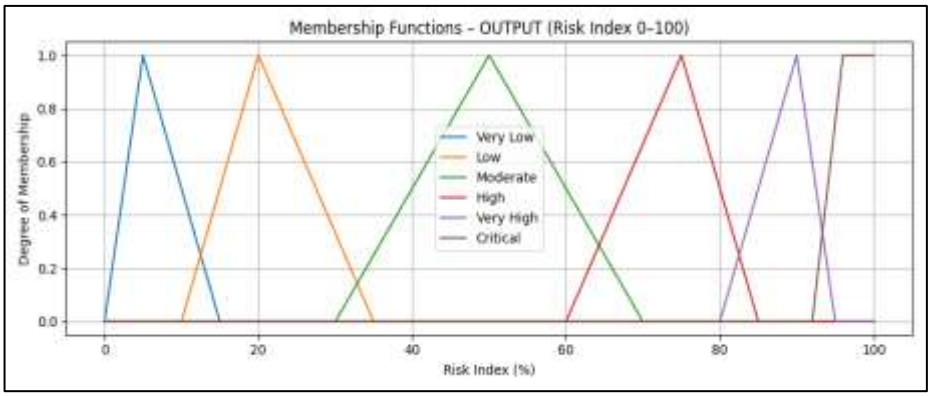


Figure 14: Output membership functions.
Source: Authors, (2026).

IV.1.1 How the fuzzy system works

As discussed earlier in topic IV, the operation of the developed fuzzy system was tested with simulated PPM input values, reflecting real industrial process values. The simulated values were a vector with the values [20, 150, 400, 800, 2500, 6000, 11000], thus covering all possible scenarios of an industrial process. Below, all tests with the aforementioned input values will be presented.

a) For PPM = 20

In Figure 15, it can be observed that the value PPM = 20 belongs almost entirely to the "Excellent" set, showing a high degree of membership. The other sets ("Good", "Acceptable", "Alert", "Bad", and "Critical") have zero membership, indicating that the process is exclusively framed within the best quality condition.

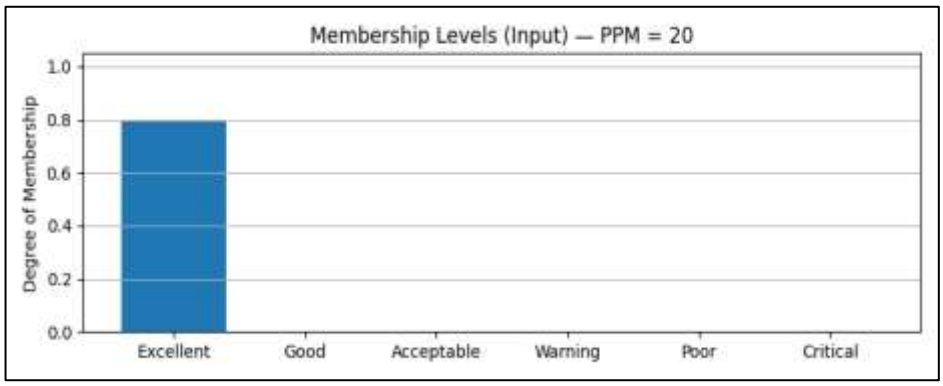


Figure 15: Degree of membership for PPP = 20.
Source: Authors, (2026).

In Figure 16, only the rule associated with the "Excellent" set is activated. The strength of this rule is equivalent to the degree of relevance calculated in the input, demonstrating that the system exclusively uses knowledge related to high-performing processes to generate its output. No other rule contributes to the final calculation, reinforcing the natural classification of the process as low-risk.

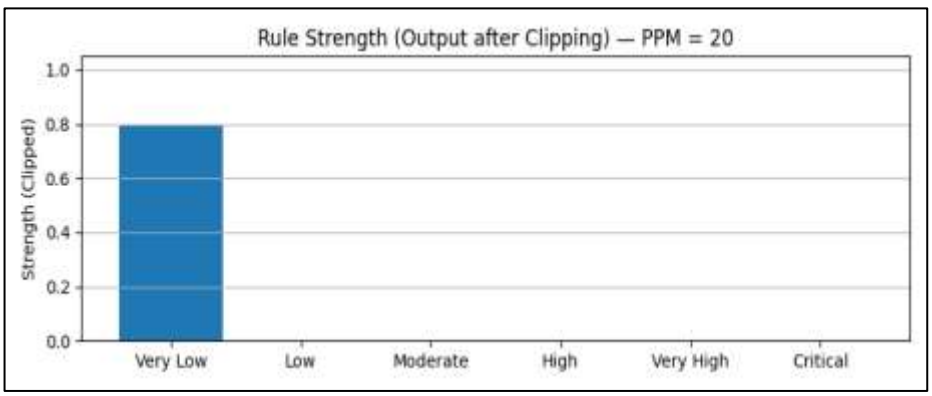


Figure 16: Rule strength for PPP = 20.
Source: Authors, (2026).

In figure 17, the graph shows that the output surface is predominantly shaped by the "Very Low Risk" set, resulting in an aggregate curve concentrated at the beginning of the risk axis. The dashed line indicates the defuzzified value, which approaches a small value, representing an extremely low risk index.

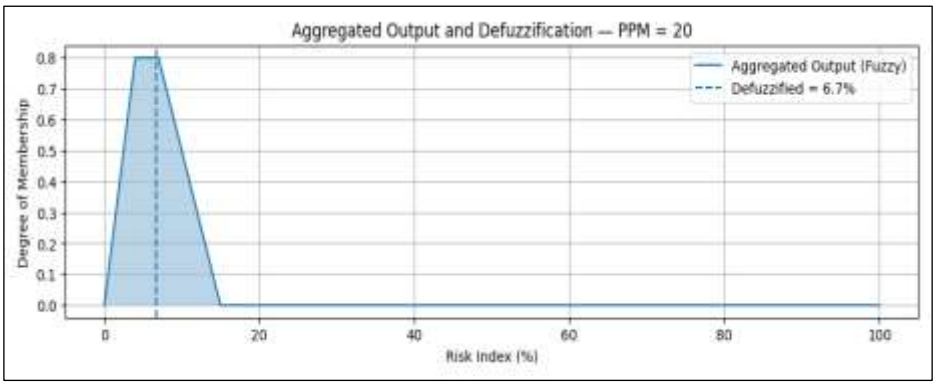


Figure 17: Aggregate output for PPP = 20.
Source: Authors, (2026).

b) For PPM = 400

In Figure 18, it can be observed that the value PPM = 400 belongs to the "Good" set, presenting a membership degree of 0.50. The other sets ("Excellent", "Acceptable", "Alert", "Poor", and "Critical") have zero membership, indicating that the process is exclusively framed within the best quality condition.

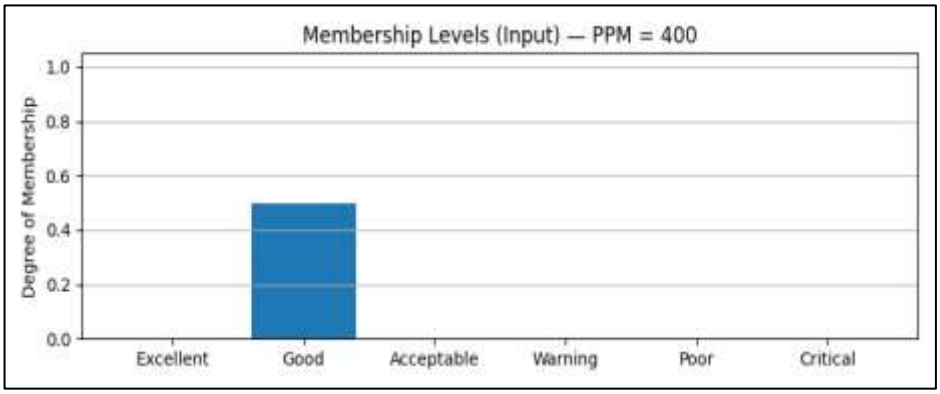


Figure 18: Membership degree for PPP = 400.
Source: Authors, (2026).

In Figure 19, only the rule associated with the "Low" set is activated. The strength of this rule is equivalent to the degree of membership calculated in the input, which demonstrates that the system exclusively uses knowledge related to high-performance processes to generate its output. No other rule contributes to the final calculation, reinforcing the natural classification of the process as low risk.

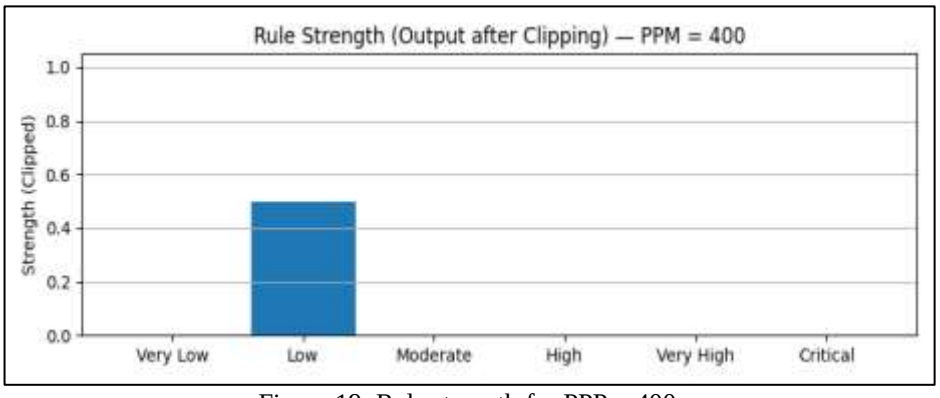


Figure 19: Rule strength for PPP = 400.
Source: Authors, (2026).

In figure 20, the graph shows that the output surface is predominantly shaped by the “Low Risk” set. The dashed line indicates the defuzzified value, which approaches a small value, representing an extremely low risk index.

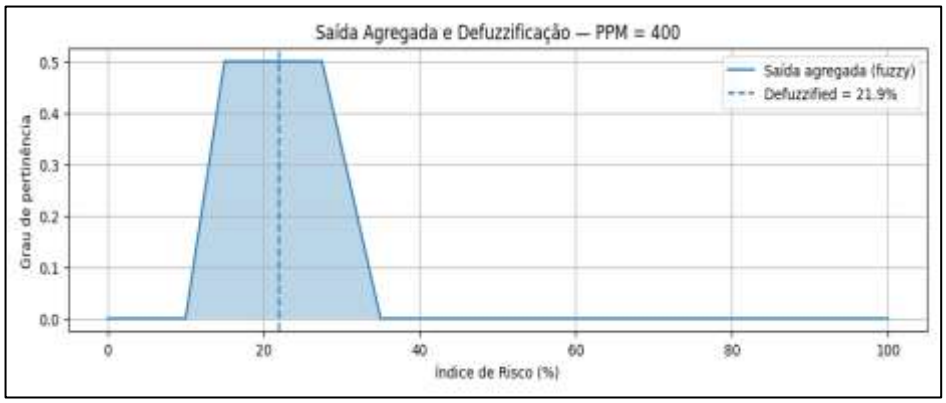


Figure 20: Aggregate output for PPP = 400.
Source: Authors, (2026).

c) For PPM = 800

In Figure 21, it can be observed that the value PPM = 800 belongs entirely to the "Acceptable" set, presenting a membership degree of 0.80. The other sets ("Excellent", "Good", "Alert", "Bad" and "Critical") have zero membership, indicating that the process is exclusively classified as acceptable.

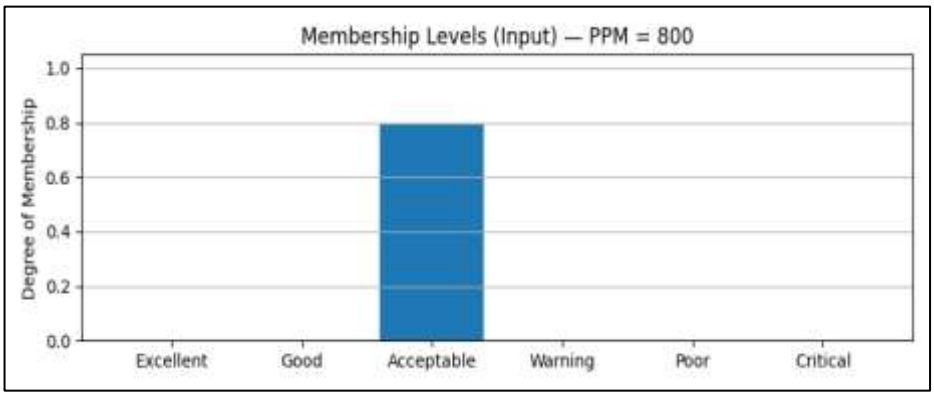


Figure 21: Membership degree for PPP = 800.
Source: Authors, (2026).

In Figure 22, only the rule associated with the "Moderate" set is activated. The strength of this rule is equivalent to the degree of relevance calculated in the input, which demonstrates that the system exclusively uses knowledge related to high-performance processes to generate its output. No other rule contributes to the final calculation, reinforcing the natural classification of the process as moderate risk.

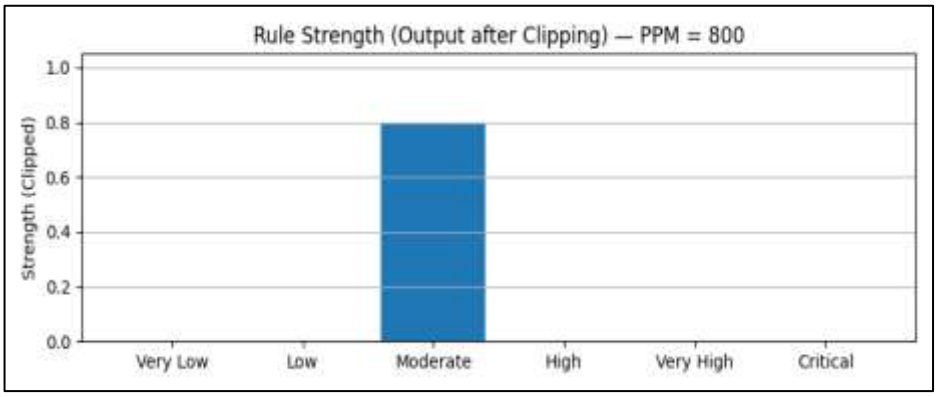


Figure 22: Rule strength for PPP = 800.
Source: Authors, (2026).

In figure 23, the graph shows that the output surface is predominantly shaped by the “Moderate” set. The dashed line indicates the defuzzified value. The system interprets that the process is within expectations, but with clear warning signs.

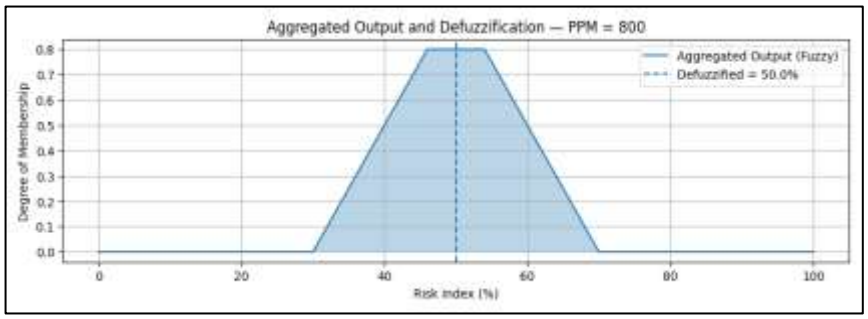


Figure 23: Aggregate output for PPP = 800.
Source: Authors, (2026).

d) For PPM = 2500

In Figure 24, it can be observed that PPM = 2500 belongs mostly to the “Alert” set, with a membership degree of approximately 0.75. The other classes (“Excellent”, “Good”, “Acceptable”, “Bad”, “Critical”) have a membership close to zero.

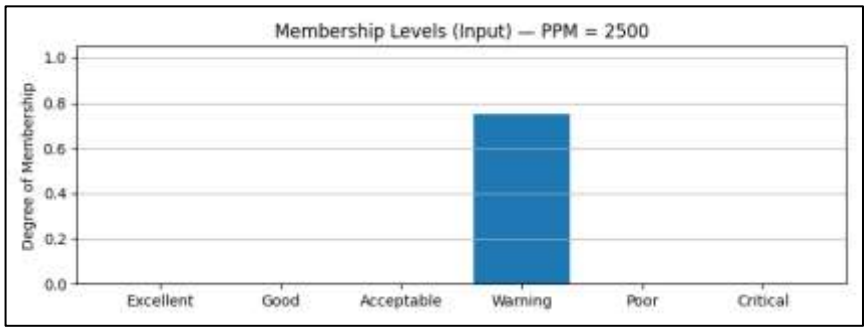


Figure 24: Membership degree for PPM = 2500.
Source: Authors, (2026).

In Figure 25, only the rule associated with the "High" set is activated. The strength of this rule is equivalent to the degree of membership calculated in the input, which demonstrates that the system exclusively uses knowledge related to high-performing processes to generate its output. No other rule contributes to the final calculation, reinforcing the natural classification of the process as High risk.

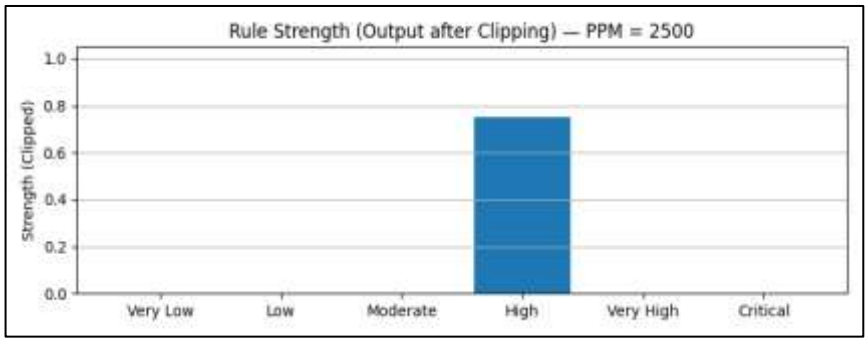


Figure 25: Rule strength for PPP = 2500.
Source: Authors, (2026).

In Figure 26, the graph shows that the output surface is predominantly shaped by the "High" set. The dashed line indicates the defuzzified value. The system interprets this as the process being in a high alert/risk zone.

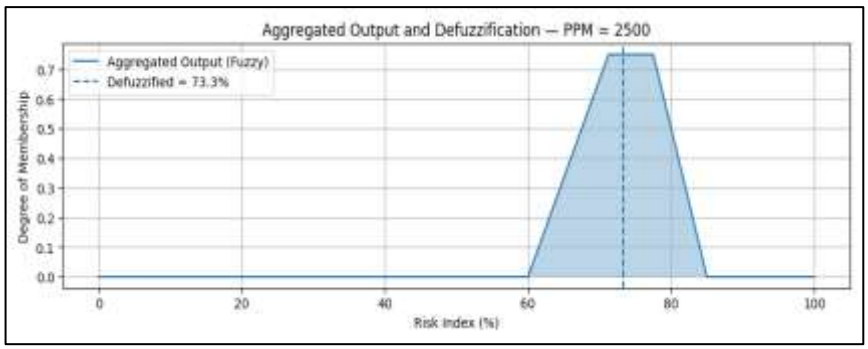


Figure 26: Aggregate output for PPP = 2500.
Source: Authors, (2026).

e) For PPM = 6000

In Figure 27, it can be observed that PPM = 6000 belongs mostly to the "Poor" set, with a membership degree of approximately 0.40. The other classes ("Excellent", "Good", "Acceptable", "Alert", "Critical") have zero membership. The system considers that PPM = 6000 is clearly a value outside the acceptable level, but it is not yet extreme.

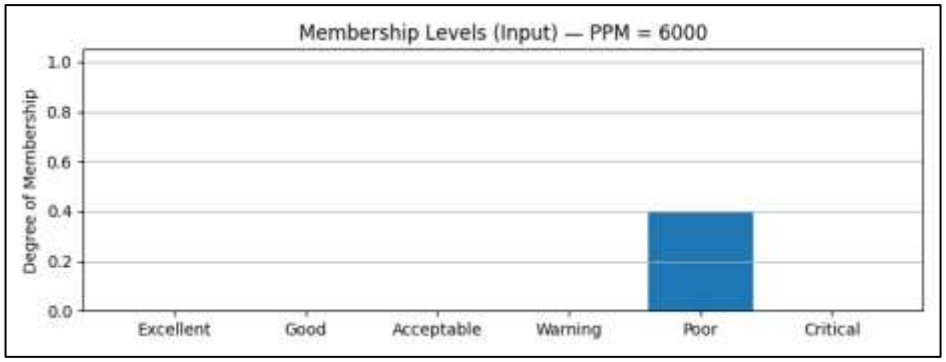


Figure 27: Membership degree for PPM = 6000.
Source: Authors, (2026).

In figure 28, only the rule associated with the "Very High" set is activated. The strength of this rule is equivalent to the degree of relevance calculated in the input, which demonstrates that the system exclusively uses knowledge related to high-performing processes to generate its output. No other rule contributes to the final calculation, reinforcing the natural classification of the process as Very High risk.

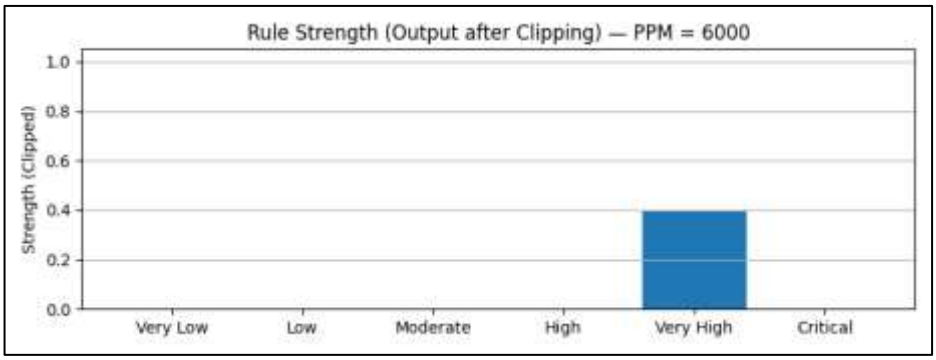


Figure 28: Rule strength for PPP = 6000.
Source: Authors, (2026).

In figure 29, the graph shows that the output surface is predominantly shaped by the "Very High" set. The dashed line indicates the defun zzified value. The system interprets this as the process being in a very high risk zone.

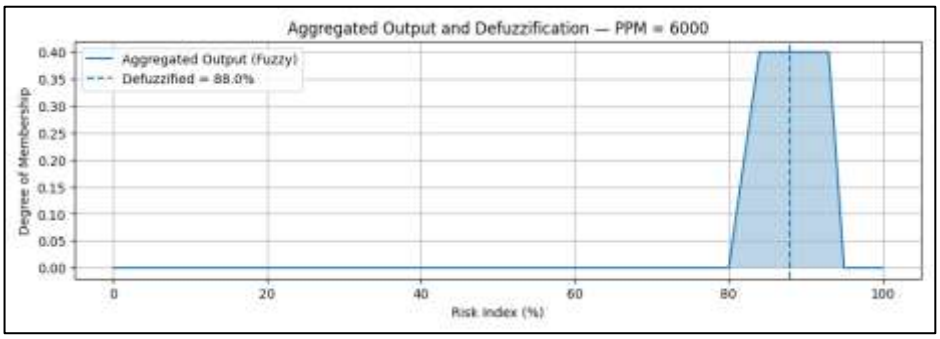


Figure 29: Aggregate output for PPP = 6000.
Source: Authors, (2026).

a) For PPM = 11000

In Figure 30, it can be observed that PPM = 11000 belongs entirely to the "Critical" set, with a membership degree of approximately 1.0. The other classes ("Excellent", "Good", "Acceptable", "Bad", "Critical") have zero membership. There is no ambiguity: the entry is at the maximum plateau of the critical set.

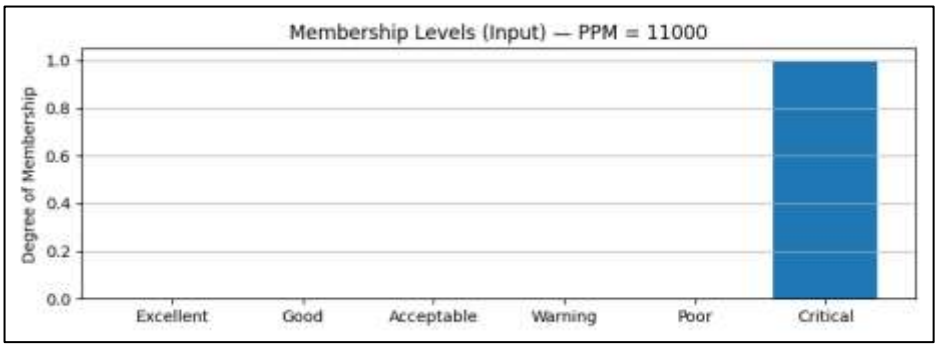


Figure 30: Membership degree for PPM = 11000.
Source: Authors, (2026).

In Figure 31, only the rule associated with the "Critical" set is activated. The strength of this rule is equivalent to the degree of relevance calculated in the input, demonstrating that the system exclusively uses knowledge related to high-performance processes to generate its output. No other rule contributes to the final calculation, reinforcing the natural classification of the process as Critical risk.

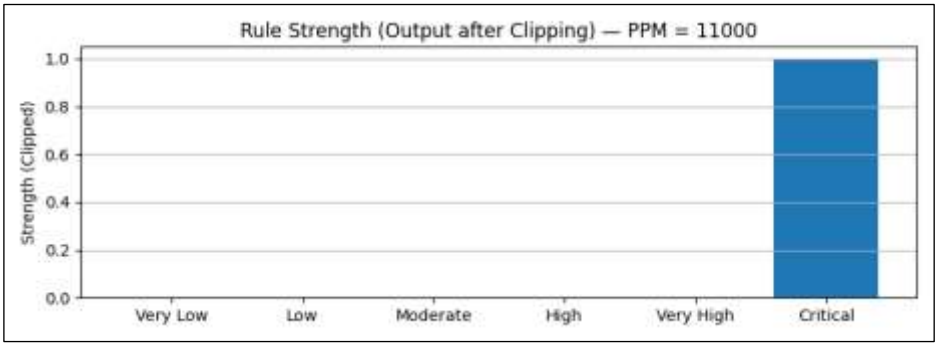


Figure 31: Rule strength for PPP = 11000.
Source: Authors, (2026).

In Figure 32, the graph shows that the output surface is predominantly shaped by the "Critical" set. The dashed line indicates the defuzzified value. The system interprets that the process is in the Critical risk zone.

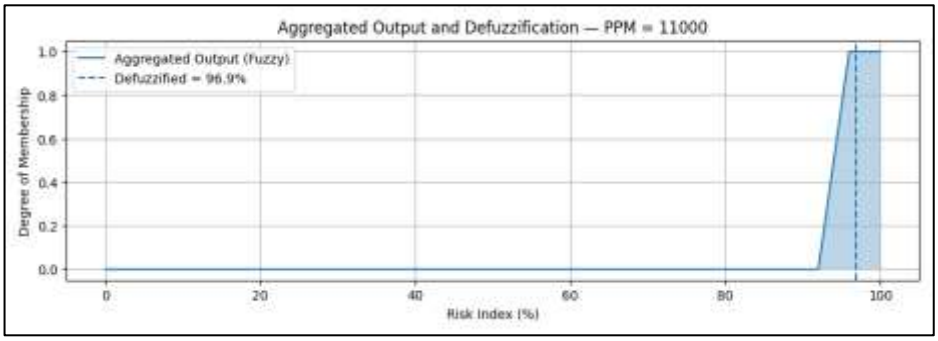


Figure 32: Aggregate output for PPP = 11000.
Source: Authors, (2026).

IV.1.2 Results and discussion

According to the analysis of the results, the developed fuzzy system was able to:

a) Faithfully represent the expected behavior of industrial processes.

- Extremely low values (such as PPM = 20) were correctly classified as a stable situation, aligning with a robust process, with minimal variability and adequate control.For intermediate values, such as 150 and 400 PPM, the
- system does not present binary responses, but rather gradual degrees of risk, which more realistically represent the uncertainty and the transition between stability and attention.

b) Provide sensitivity to the process, even in "gray" regions.
In the real industrial scenario, a process with 400 to 800 PPM may not be considered completely out of control, but it certainly requires monitoring. The fuzzy system managed to capture this, classifying them as:

- 400 PPM → moderate risk
- 800 PPM → high risk

This differentiation is essential to avoid both false alarms and excessive complacency.

c) Handling Severe Scenarios Well

In tests with high values (2500, 6000, and 11000), the system reacted appropriately, raising the belonging function in the maximum risk categories, ensuring that clearly unstable processes were classified as "critical." This behavior reflects industrial logic, in which PPM above 2000 generally indicate:

– process out of control;
– raw material, setup, or calibration problems;
– systemic failures or equipment wear.

In environments where standards such as IATF 16949 or ISO 9001 are applied, processes with very low PPM indicate maturity and stability, while high values require immediate corrective actions. The fuzzy system demonstrated the ability to simulate this logic, but with greater flexibility, because:

a) it does not depend exclusively on fixed limits;
b) it incorporates human knowledge through linguistic rules;
c) it allows smooth transitions between risk ranges.

This means that the fuzzy model can behave like an expert, evaluating real process conditions. The tests performed demonstrated that the fuzzy system is capable of detecting increasing levels of non-conformity in a coherent, continuous manner, aligned with the real needs of industrial manufacturing. The interpretation of each PPM range confirmed that the system:

a) reproduces the behavior of a quality expert,
b) provides gradual and reliable diagnoses,
c) assists in operational and strategic decision-making,
d) can replace or complement traditional SPC systems.

Below is a summary of the tested values, aggregated outputs, and recommended action by the fuzzy logic:

Table 2: Summary of tested values, aggregated outputs, and action recommended by fuzzy logic.

PPM	Active Relevance Set	Fuzzy System Interpretation	Resulting Risk Index	Recommended Action in the Industry
20	Excellent (≈1.0)	Process in excellent condition.	Very Low Risk	No action required.
150	Good (≈0.50), Excellent (≈0.00)	Partially belonging to the "Good" group.	Low Risk	Perform periodic checks.
400	Acceptable (≈1.0)	The process was fully classified as "Acceptable".	Moderate Risk	Monitor trends; initiate corrective actions.
800	Alert (≈1.0)	Process reaches alert level. System identifies high risk.	High Risk	Review machine, process parameters, calibration, and training. Initiate action plan.
2500	Poor (≈1.0)	Process clearly inadequate. Strong activation of the “Poor” category.	Very High Risk	Stop the process for corrections; perform maintenance.
6000	Poor (≈1.0) → near critical	The process is still within the "Bad" category, but close to the "Critical" threshold.	Very High Risk / Trending Towards Critical	Immediate action: Block the batch, conduct root cause analysis, and implement urgent correction.
11.000	Critical (≈1.0)	Full relevance in "Critical". System indicates maximum severity.	Extreme Risk	Complete shutdown; process audit; emergency actions; rework.

Source: Authors, (2026).

Thus, the application of the developed system demonstrates that fuzzy logic is a powerful and suitable tool for implementing intelligent non-conformity detection systems.

V. CONCLUSIONS

This dissertation presented the development and validation of a Non-Conformance Detection System for Industrial Manufacturing Processes based on Fuzzy Logic, capable of analyzing process performance using the PPM (Parts Per Million) indicator and providing intelligent diagnoses of the risk level associated with product quality. The results obtained demonstrated that the fuzzy system is capable of interpreting input values continuously, flexibly, and in line with industrial reality, providing understandable classifications such as low, moderate, high, critical risk, or widespread failure. The fuzzy approach proved particularly useful for processes where binary interpretation (accept/reject) is insufficient, allowing the system to represent uncertainties, gradual transitions, and intermediate scenarios with greater sensitivity. The development of the fuzzy system, with simulation in various scenarios, fulfills the general and specific objectives that comprise this dissertation, as it demonstrated excellent capacity to:

a) Correctly classify all tested PPM ranges (20, 150, 400, 800, 2500, 6000 and 11000), reflecting the expected behavior of an industrial process under different compliance conditions;
b) Provide continuous and realistic responses, avoiding abrupt changes common in traditional deterministic systems;
c) Represent expert knowledge through fuzzy rules and linguistic terms;
d) Assist in industrial decision-making, intuitively identifying situations ranging from full stability to systemic failure.

This behavior confirms that the system is suitable for manufacturing monitoring, internal audits, improvement programs, and operational control.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Roberto Ferreira de Lima and Helder Kiyoshi Miyagawa.
Methodology: Roberto Ferreira de Lima, Helder Kiyoshi Miyagawa and Rivanildo Duarte Almeida.
Investigation: Roberto Ferreira de Lima and Helder Kiyoshi Miyagawa.
Discussion of results: Roberto Ferreira de Lima and Helder Kiyoshi Miyagawa.
Writing – Roberto Ferreira de Lima.
Writing – Review and Editing: Roberto Ferreira de Lima and Rivanildo Duarte Almeida.
Resources: Roberto Ferreira de Lima.
Supervision: Helder Kiyoshi Miyagawa.
Approval of the final text: Helder Kiyoshi Miyagawa, Roberto Ferreira de Lima and Rivanildo Duarte Almeida.

VII. REFERENCES

[1] Slack, N., Brandon-Jones, A., Burgess, N. (2020). Operations Management. Pearson.

[2] Montgomery, D. C. (2019). Introduction to Statistical Quality Control. Wiley.

[3] Zadeh, L. A. (1996). Fuzzy Logic = Computing with Words. IEEE Transactions on Fuzzy Systems.

[4] Kasabov, N. (2020). Time-Space, Spiking Neural Networks and Brain-Inspired Artificial Intelligence. Springer.

[5] Schumacher, Simon, et al. Conceptualization of a Framework for the Design of Production Systems and Industrial Workplaces. 30th CIRP Design 2020 (CIRP Design 2020). Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Fraunhofer Institute for Industrial Engineering IAO, Institute of Industrial Manufacturing and Management, University of Stuttgart, Allmandring, Stuttgart, Germany. 2020.

[6] Wagner, Tobias, Herrmann, Christoph, Thied, Sebastian. Industry 4.0 impacts on lean production systems. Chair of Sustainable Manufacturing & Life Cycle Engineering, Institute of Machine Tools and Production Technology (IWF), Technische Universität Braunschweig, 2017.

[7] Rahman, A. Azwan Abdul. "Revolution of Production System for the Industry 4.0". IntechOpen, Mar. 11, 2020. doi: 10.5772/intechopen.90772. CAPÍTULO 1, 2020.

[8] Dombrowski, Uwe, Richter, Thomas, Krenkel, Philipp. Interdependencies of industrie 4.0 & Lean Production Systems - a use cases analysis -. Institute for Advanced Industrial Management, Technische Universität Braunschweig, 2017.

[9] Bihi, T., Luwes, N., Kusakana, K. Innovative Quality Management System for Flexible Manufacturing Systems. Department of Electrical, Electronic and Computer Engineering, Central University of Technology Bloemfontein, South Africa, 2018.

[10] Akdogan, Anil, Vanli, Ali Serdar. 'Introductory Chapter: Mass Production and Industry 4.0', Mass Production Processes. IntechOpen, doi: 10.5772/intechopen.90874, 2020.

[11] Zhong, Ray Y., et al. Intelligent Manufacturing in the Context of Industry 4.0: A Review. Department of Mechanical Engineering, The University of Auckland, Auckland 1142, New Zealand, Industry 4.0 Campaign, Festo AG & Co. KG, Esslingen 73726, Germany, Department of Mechanical Engineering, University of Bath, Bath BA2 7AY, UK, 2017.

[12] Kusiak, Andrew. Smart manufacturing. Intelligent Systems Laboratory, Department of Mechanical and Industrial Engineering, The University of Iowa, Iowa City, IA, USA, 2018.

[13] Li, Bo-hu et al. Applications of artificial intelligence in intelligent manufacturing: a review. The Second Academy of China Aerospace Science and Technology Corporation, Beijing 100039, China, Beijing Aerospace Intelligent Manufacturing Technology Development Co., Ltd., Beijing 100039, China, 2017.

[14] Lenz, Juergen et al. Optimizing smart manufacturing systems by extending the smart products paradigm to the beginning of life. Industrial and Management Systems Engineering, Benjamin M. Statler College of Engineering and Mineral Resource, West Virginia University, Morgantown, WV 26506, United States, Advanced Manufacturing Research Center, Youngstown State University, Youngstown, OH 44555, United States, McNair Aerospace Center, University of South Carolina, Columbia, SC 29208, United States, 2020.

[15] Carneiro, Flávio Araújo. *Sistema de gestão da qualidade: uma revisão bibliográfica*. Research, Society and Development, v. 9, n. 10, p. e4309108373, 2020. Disponível em: https://rsdjournal.org/index.php/rsd/article/download/43006/34685/454384. Acesso em: 13 ago. 2025.

[16] Lima, Sávio Araújo; Seleme, Rogério. Qualidade 4.0: uma revisão sobre os impactos da indústria 4.0 na gestão da qualidade. In: Congresso Brasileiro de Engenharia de Produção – CONBREPRO, 10., 2021, Ponta Grossa. Anais [...]. Ponta Grossa: APREPRO, 2021. Disponível em: https://aprepro.org.br/conbrepro/2021/anais/arquivos/09272021_210901_615266f53f133.pdf. Acesso em: 13 ago. 2025.

[17] Hoffmann, Vanderlei. Gestão da qualidade na Indústria 4.0. Ponta Grossa: Universidade Tecnológica Federal do Paraná (UTFPR), 2019. Disponível em: https://repositorio.utfpr.edu.br/jspui/bitstream/1/25908/1/PB_ESEP_IV_2019_21.pdf. Acesso em: 13 ago. 2025.

[18] Moraes, Simone de Sá de. Implementação dos requisitos da ISO 9001:2015 em vidrarias industriais. Universidade do Minho. Dissertação (Mestrado em Engenharia Industrial) – Universidade do Minho, Braga, 2023. Disponível em: https://repositorium.sdum.uminho.pt/handle/1822/84142. Acesso em: 13 ago. 2025.

[19] Peixoto, Rafaela Aparecida Martins. Gestão da qualidade na indústria: estudo de caso. Dissertação (Mestrado em Engenharia de Produção) – Universidade Presidente Antônio Carlos, Barbacena, 2013. Disponível em: https://ri.unipac.br/repositorio/wp-content/uploads/tainacan-items/282/283918/RAFAELA-AP.-Martins-Peixoto-Gestao-da-Qualidade-na-Industria-Engenharia-de-Producao-2013.pdf. Acesso em: 13 ago. 2025.

[20] Wiley, John. Statistical Quality Control: Using Minitab, R, JMP, and Python. Phase I Quality Control Charts for Variables, 2021.

[21] Zinchenko, Timur O. Development of a Quality Control System for Transparent Conductive Oxides, Penza State University, Russia, 2020.

[22] Ciecińska, Barbara (2023). Identification of Defects Causes: Ishikawa Diagram and 5 Whys in Theoretical and Practical Terms. Industrial Engineering and Management. IntechOpen, May 29, 2024. doi: 10.5772/intechopen.113990.

[23] Chaves, M.L., et al. Experimental assessment of quality in injection parts using a fuzzy system with adaptive membership functions. Department of Computer and Systems Engineering, Universidad Católica de Colombia, Caracas, Bogotá - Colombia, Departmento de Ingenierías Mecánica, informática y Aeroespacial, Universidad de León, León, Spain, Departamento Ingeniería Mecánica. Universidad de La Frontera. Temuco, Chile, Departamento de Ingeniería Mecánica. Universidad Politécnica de Madrid, José Gutiérrez Abascal, Madrid, Spain, 2019.

[24] Amin, S. H., Razmi, J., and Zhang, G. Supplier selection and order allocation based on fuzzy SWOT analysis and fuzzy linear programming, Expert Systems with Applications, vol. 38, pp. 334-342, 2011.

[25] Almeida, Rivanildo Duarte. Análise de Impactos Harmônicos em Rede de Distribuição de Energia Elétrica de Média Tensão Utilizando Técnicas de Inteligência Computacional. Universidade Federal do Pará. Instituto de Tecnologia. Programa de Pós-Graduação em Engenharia de Processos – PPGEP, Belém, UFPA, 2018.

[26] Aquino, Dener Jeferson Horta de. Análise de impactos harmônicos em redes elétricas: um estudo comparativo entre as técnicas regressão linear e árvore de decisão. Universidade Federal do Pará – UFPA, 2022.

[27] Khadeer, Ahmed. Brain-Inspired Spiking Neural Networks. DOI: 10.5772/intechopen.93435, 2020.

[28] HERNÁNDEZ-HERNÁNDEZ, Marisol, CRUZ, Luis Alfonso Bonilla e COBIÁN-ROMERO, Lizbeth. Improvement of Validated Manufacturing Processes with Fuzzy Logic. INTECHOPEN, DOI: 10.5772/intechopen.113302, 2023.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 483-495. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3122

RESEARCH ARTICLE OPEN ACCESS

INTELLIGENT MODEL FOR PRODUCTIVE PLANNING OF ELECTRONIC METERS WITH INDUSTRY 4.0 TECHNOLOGICAL RESOURCES

Jandecy Cabral Leite[1], Marcelo Maia do Nascimento[2], Marivan Silva Gomes[3], Alan Ferreira Pinheiro Tavares[4], André Barbosa de Lima[5], Matheus Pereira de Souza[6], Daniel Gomes Gulart[7], Marcos Castro de Menezes[8]

[1, 2,3,4,5,6]Department of Industrial Research Department of the Galileo Institute of Technology and Education of Amazonia (ITEGAM). Address: Joaquim Nabuco, No. 1950. Neighborhood: Downtown. ZIP CODE: 69020-030. Manaus-Amazonas, Brazil.
[7,8] Wasion da Amazônia Electronic Instruments Industry LTDA. Av. Cosme Ferreira, No. 12.110. Colônia Antônio Aleixo. ZIP CODE: 69008-310. Manaus – Amazonas, Brazil.

[1]https://orcid.org/0000-0002-1337-3549, [2]http://orcid.org/0000-0001-7183-8642, [3]http://orcid.org/0009-0007-2022-9047
[4]http://orcid.org/ 0000-0003-1644-9999, [5]http://orcid.org/0009-0002-6910-0635, [6]http://orcid.org/0009-0008-3447-3566
[7]http://orcid.org/0009-0002-2787-7753, [8]http://orcid.org/0009-0000-7874-9132

Email: *jandecy.cabral@itegam.org.br, marcelo.nascimento@itegam.org.br, msgomes@uea.edu.br, alan.tavares@itegam.org.br, amilandre@gmail.com, matheus@itegam.org.br, daniel.gulart@wasion.com.br, marcos.castro.menezes@gmail.com

ARTICLE INFO

Article History
Received: December 12, 2025
Revised: December 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Industry 4.0, Artificial Intelligence, Production Planning, Electronic Meters, Operational Optimization.

ABSTRACT

The advancement of digital technologies has significantly transformed industrial processes, shaping the Industry 4.0 scenario, in which cyber-physical systems, IoT, Big Data, and Artificial Intelligence (AI) enhance productivity efficiency and flexibility. In this context, the manufacturing of electronic meters faces challenges such as operational complexity, demand variability, the need for traceability, and data integration across the production chain. Therefore, the adoption of intelligent solutions that support production planning and control in an agile and precise manner becomes essential.

This study aims to develop an intelligent model for the productive planning of electronic meters, integrating Industry 4.0 resources to improve decision-making and optimize operational flow. The model seeks to minimize bottlenecks, forecast material requirements, improve machine utilization, and increase planning reliability. The methodology was structured in four stages: mapping the production process and identifying critical variables; real-time data integration via IoT and MES systems; development of AI models based on machine learning; and the construction of a system capable of simulating scenarios and generating production plans. Tests using historical and real data were conducted. The results indicated higher accuracy in forecasts, reduced idle times, and better balance between capacity and demand. The model proved adaptable to different scenarios, demonstrating its potential to support companies in adopting more dynamic and efficient processes aligned with Industry 4.0.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Industry 4.0 has established itself as one of the pillars of digital transformation in the industrial environment, driving integration between advanced technologies, real-time data, and intelligent production processes [1-3]. Characterized by the use of cyber-physical systems, connectivity via the Internet of Things (IoT), Big Data, advanced automation, and Artificial Intelligence (AI), this new industrial era provides significant gains in efficiency, flexibility, and analytical capacity [4-7]. In this scenario, the competitiveness of companies is directly related to the adoption of tools capable of improving decision-making and promoting Operational Optimization at all levels of production.

One of the sectors that most demands such advances is the manufacture of electronic meters, whose production process involves a wide variety of components, high precision requirements, and a constant need for traceability [8-10]. Production planning in this type of environment presents complex challenges, marked by demand variation, diversity of meter models, and the need to integrate information from different stages of the production chain. Traditional planning strategies, based on fixed rules or manual analysis, are limited in the face of growing operational complexity [11],[12]. Thus, there is a need for intelligent and adaptive solutions capable of dealing with dynamic and variable environments. The application of AI in this context offers significant opportunities for analyzing large volumes of data, forecasting scenarios, identifying bottlenecks, and supporting real-time decision-making, thereby increasing productivity and quality levels [13-15]. The proposal to develop an intelligent model for the production planning of electronic meters represents a significant contribution to the field, as it integrates essential Industry 4.0 technologies with advanced AI methods geared toward optimization. The innovation of this model lies in its ability to combine data collected by IoT sensors, Manufacturing Execution Systems (MES), and analytics platforms to create a more autonomous, responsive, and efficient production environment.

In addition, the use of machine learning algorithms enables more accurate predictions about production times, future demand, resource availability, and operational performance, resulting in more reliable production plans that are aligned with market needs [16-18]. The relevance of this topic is reinforced by the competitive pressures faced by the industrial sector, which demands rapid customer response, cost reduction, and increased overall efficiency. By incorporating elements of Industry 4.0 into the planning process, companies now have tools capable of transforming raw data into strategic information, promoting more assertive decisions and reducing uncertainty. The proposed model also contributes to improving traceability, increasing process reliability, and reducing waste, which are fundamental aspects for lean and sustainable production [19], [20]. Thus, this work presents an innovative approach that integrates intelligent planning, digital technologies, and Operational Optimization, demonstrating how Industry 4.0 can enhance productive performance and strengthen competitiveness in the electronic meter sector.

II. THEORETICAL REFERENCE

II.1 APPROACH TO INDUSTRY 4.0 AND ITS DEGREES OF MATURITY

Industry 4.0 represents a milestone in the evolution of production systems, characterized by the integration of advanced digital technologies, intelligent automation, and massive data analysis [21-23]. For manufacturers of electronic meters, whose production requires a high degree of precision, traceability, and reliability, the adoption of these concepts becomes not only strategic but fundamental to ensuring competitiveness and quality in an increasingly regulated and demanding market [24]. The application of Industry 4.0 in this sector generally occurs gradually, following different degrees of maturity. The first stage, known as initial digitization, involves replacing manual processes with computerized systems. For electronic meter manufacturers, this may include the digitization of production orders, electrical test records, calibration, and quality inspections. Although there is still no significant integration between systems, this stage lays the foundation for further advances, reducing errors and improving the traceability of components and batches [25]. In the second stage of maturity, data connectivity and integration emerge.

Here, welding machines, SMT (Surface Mount Technology) devices, test benches, and optical inspection systems are integrated into an MES (Manufacturing Execution System) [26]. This connection allows production data—such as cycle times, failures, welding parameters, and electrical test results—to be collected automatically and analyzed in real time. At an electronic meter manufacturer, this stage enables rapid detection of deviations, reducing rework rates and ensuring greater consistency in the performance of the equipment produced. The third stage corresponds to intelligent automation and the use of advanced technologies such as artificial intelligence, machine learning, and cyber-physical systems. At this level, electronic meter assembly lines become capable of automatically adjusting machine parameters according to observed conditions. For example, algorithms can predict failures in welding equipment, anticipating maintenance needs. In addition, computer vision systems can enhance the inspection of electronic boards, identifying micro-defects that would go unnoticed by the human eye. The analysis of historical data also makes it possible to predict variations in energy consumption in the meters, optimizing the final calibration.

At the highest level of maturity, the factory reaches an autonomous and fully integrated stage, in which production flows, internal logistics, inventory management, and quality control operate in a synchronized manner. Digital twins — digital replicas of production processes — allow scenario simulations, testing of new meter models, and validation of production parameters even before physical manufacturing. Continuous communication between ERP, MES, and IoT devices ensures a highly efficient and adaptable operation [16],[17]. For the electronic meter manufacturer, this results in shorter response times, significant cost reductions, and greater reliability in measurements and calibrations. Thus, the evolution through the maturity levels of Industry 4.0 is not merely a technological modernization, but a profound transformation of production processes. In companies that manufacture electronic meters, this journey represents the key to elevating quality standards, increasing competitiveness, and meeting the demands of a sector driven by precision, safety, and continuous innovation.

II.2 PRODUCTION PLANNING AND CONTROL FOR ELECTRONIC METERS

Production Planning and Control (PCP) plays a strategic role in the electronic meter industry, a segment that demands high precision, regulatory compliance, and complete product traceability. The complexity involved in manufacturing these devices—which include electronic boards, sensors, power supplies, communication modules, and calibration systems—makes PCP essential to ensuring operational efficiency, final quality, and on-time delivery [27-30]. The PCP process begins with long-term planning, which involves demand forecasting, production capacity analysis, and the definition of supply strategies. For electronic meter manufacturers, this stage is particularly sensitive, as demand fluctuations can occur due to public tenders, large utility projects, or technology replacement cycles.

Good planning must consider these variables, adjusting production capacity, the allocation of skilled labor, and the availability of critical inputs, such as microcontrollers and communication components, which are subject to variations in the global market. As part of medium-term planning, PCP transforms forecasts into master production plans, defining which meter models will be produced in each period. This includes balancing assembly lines, defining economic batches, and sequencing operations. Since many electronic meters have variations in communication modules (such as Wi-Fi, PLC, or RF) and advanced features, PCP must optimize production to minimize setup changes and maintain process fluidity. The operational stage, or production control, involves daily monitoring of the shop floor. At this point, the PCP monitors indices such as SMT line efficiency, soldering machine availability, functional test results, and rework rates. Integrated systems, such as MES and ERP, are widely used to record production data in real time, enabling quick decisions when deviations arise. For example, if a batch of boards shows recurring failures in electrical tests, the PCP can intervene by adjusting process parameters, reallocating resources, or engaging product engineering [16-18].

Another essential aspect is materials control, since the production of electronic meters depends on specific components that are often difficult to replace on the international market. PCP must ensure adequate stock levels, avoiding interruptions in production and preventing excess obsolete items. Techniques such as MRP (Material Requirements Planning) and ABC analysis assist in this management, ensuring that critical components receive priority [31]. Quality control is also part of PCP, as meters must comply with national and international metrological standards and regulations. Planning must allow adequate time for calibration, inspections, functional tests, and certifications, avoiding bottlenecks that compromise delivery times [32],[33]. Finally, PCP plays a crucial role in aligning production, engineering, purchasing, and logistics. In the electronic meter industry, where product reliability is essential for energy security and accurate consumption measurement, a well-structured PCP ensures not only efficiency but also business sustainability and full compliance with market requirements.

II.3 REAL-TIME DATA INTEGRATION VIA IOT

The research aims to improve this current system to encompass the development of an intelligent system for production planning, which is based on monitoring production events, which indicated actions to be taken to follow the planning carried out, integrating resource management and B.O.M, which is currently done by Jundsoft Software, optimizing business rules with new Design Patterns. As can be seen in Figure 2, the service layer has APIs developed for communication with the equipment and, therefore, continuously collects data from this equipment directly into the database.

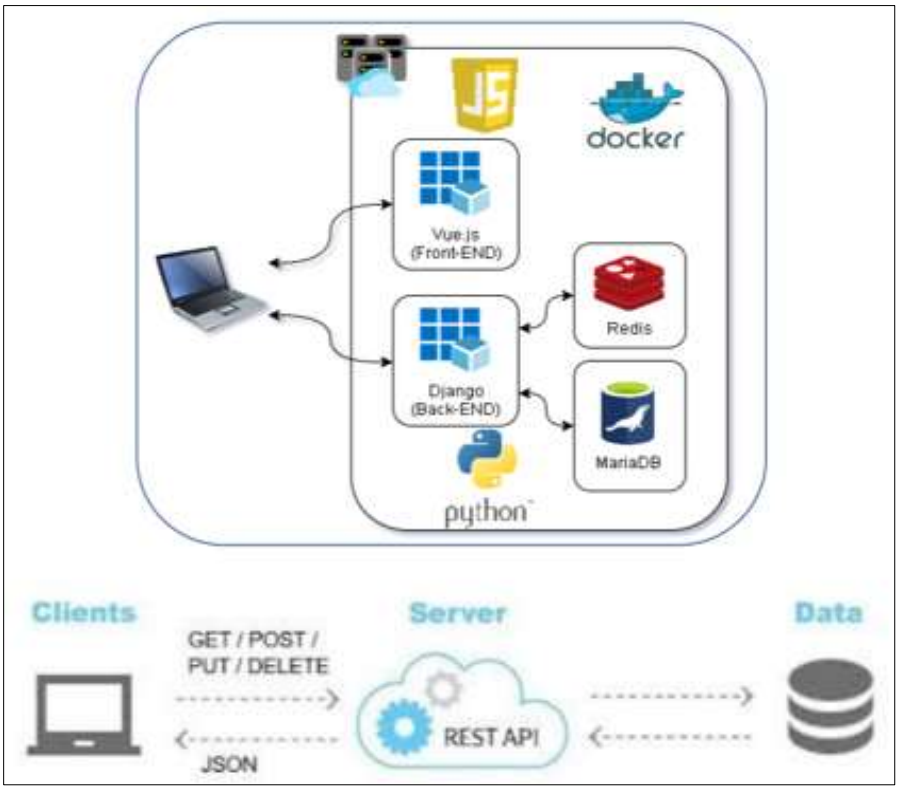


Figure 1: Draft of the Software Architecture for Intelligent MES System.
Source: Authors, (2026).

II.3.1 Traceability Device

This research aims to integrate software stacks (service layer API) with hardware stacks (Traceability Device), which in turn is connected to equipment that generally has RS232 serial communication. Therefore, the device will be the interface for communicating with the equipment and connecting them to the internal network, moving the equipment from stage 1 maturity to stage 2, where there is connectivity, for collecting data from each piece of equipment on a pilot line, which will be used to validate the collection. Some devices in this pilot production line do not have connectivity and will receive a traceability device for data collection, as shown below:

• Parameterization device:

This prototype device will be responsible for configuring the product's memory, recording the serial number, and checking other quality parameters to ensure that the product complies with the customer's purchase order requirements (verification according to the production order). This system should operate automatically, without the need for operator selection. Based on the production order number, the system will consult via API which configurations and checks should be performed on the product. As these configurations are made based on customer parameters, the name of the device is Parameterization.

• PIMA Asynchronous Serial Output Device:

This device will be a functional test prototype of the product's communication protocol. It will be responsible for ensuring that the product will function and communicate correctly, and for ensuring that the serial code is recorded in the memory (performed on the parameterization device). This step will be performed after the product is permanently closed, ensuring that there will be no cover exchange error, or that the mechanical closing process did not damage the product. To know what should be tested, the system will consult via API from the production order number what should be the correct serial number of the production sequence, which communication protocol, and other optional items to be verified.

II.4 MANUFACTURING EXECUTION SYSTEMS (MES)

Manufacturing Execution Systems (MES) play a central role in modern industrial operations management, acting as the bridge between strategic planning carried out by corporate systems, such as ERP, and operational activities on the factory floor. The main objective of an MES is to ensure that production is carried out efficiently, traceably, and in line with quality, time, and cost specifications. An MES collects real-time data from machines, operators, and processes, allowing managers to monitor production performance minute by minute. This data includes indicators such as productivity, downtime, material consumption, batch quality, and cycle time for each stage of the process. Based on this information, the company can quickly identify bottlenecks, anomalies, or deviations, making corrective decisions that increase efficiency and reduce waste. This continuous monitoring capability makes MES essential in highly technological environments, such as automated lines, electronic device assembly, and processes that require precision and repeatability.

In addition to monitoring, MES also controls production, managing manufacturing orders, guiding operators with digital instructions, and recording each step of the process, ensuring full traceability. Another relevant point is quality control, as MES integrates inspection tools, automatic collection of test results, and application of sampling plans. This ensures that non-compliant products are quickly isolated and that quality standards are strictly met [26]. In addition, MES contributes to the management of resources, including operators, machines, tools, and materials, optimizing their use. When fully integrated with ERP and automation systems, it becomes a pillar of Industry 4.0, enabling decisions based on real and reliable data. Thus, Manufacturing Execution Systems represent an indispensable tool for organizations seeking productivity, quality, and competitiveness in a constantly changing market.

III. MATERIALS AND METHODS

III.1 METHODOLOGY AND ACTION STRATEGY:

The study will be developed through research, by surveying existing data, collecting information at the factory with managers, administrators, and technicians in order to ensure the development of the project in accordance with the real needs of the company. This project will be developed by a team of developers, engineers, and technicians who will be assigned well-defined responsibilities in each phase of the project. To plan, execute, and manage this project, PMBOK principles will be adopted as shown in Figure 2. In this project, five groups of processes stand out.

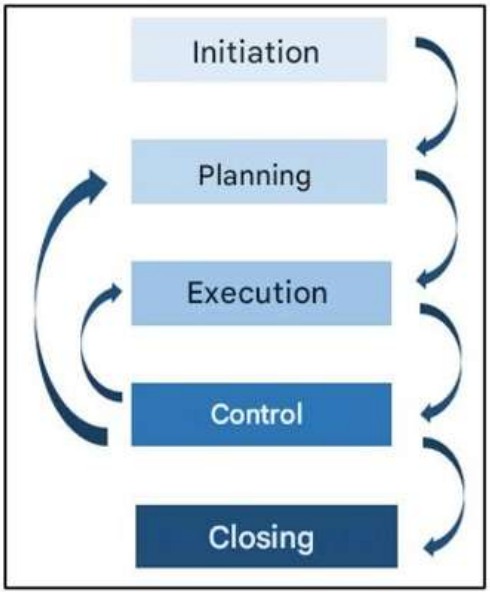


Figure 2: PMBOK management model.
Source: Authors, (2026).

• **Initiation:** In this phase, the project's needs will be listed to begin the development stages. Relevant data for the project's development will also be collected here, such as scope, risks, costs, participants, and schedule.
• **Planning:** The planning will be organized into a timeline according to the activities to be carried out. Frequent meetings are expected so the team can align ongoing work and determine ways to improve the process with agility. For this purpose, daily discussions are intended, enabling faster decision-making, value delivery to the customer, and participation from the project sponsor.
• **Execution:** The project will be executed by professionals with expertise in software, electrical, electronic, automation, and mechanical areas. This phase will also include the coordination of resources according to the project's progress and stage.
• **Control:** Project control will be performed to ensure that the objectives are achieved. This will be done through schedule analysis and meetings with team members and the project sponsor.
• **Closure:** In this stage, the client is expected to receive: reports, acceptance terms, maintenance and operation manuals (and other documents required by the Institute and the Company), as well as the new products developed. In addition, as the software development method, the incremental methodology will be used, in which each activity cycle will develop an integral part of the software, making it possible to measure the following phases of the process according to Figure 3.

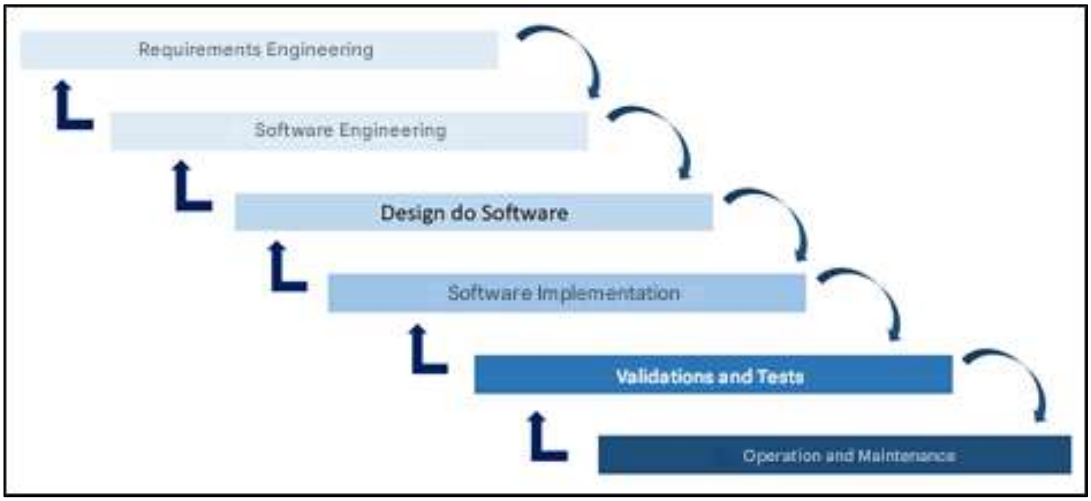


Figura 3: Ciclo incremental do Software.
Source: Authors, (2026).

• **Requirements Engineering:** In this stage, the needs related to the software—specifically the module to be developed—will be defined so that the solution meets the company's scenario, considering visual, behavioral, action-related, criteria-based, and operational limitations.
• **Software Engineering:** In this stage, diagrams and documentation regarding the module and the implemented functionalities will be developed. This document should address the software design, describing limitations, use cases, functionalities, technologies, and the technical definitions of the functional and non-functional requirements.
• **Software Design:** In this stage, the software design will be developed based on the user interaction interfaces, visual identity definitions, element layout, actions, behaviors, and company standards, taking operational limitations into account.
• **Software Implementation:** In this stage, the process of creating, defining, and executing the coding of the functionalities specified in the requirements document will be carried out, along with strategies to achieve the module to be delivered.
• **Validation and Testing:** This stage is characterized by activities that perform validations and tests to approve the application and the developed module. Functionality, integrity, security, unit, and installation tests will be executed as methods to assess product quality, support bug fixes, and carry out code refactoring processes.
• **Operation and Maintenance:** This stage will involve code refactoring and adjustments that may be necessary due to new requirements or technological updates. This includes carrying out coding activities during this phase.

III.2 TOOLS FOR PROJECT DEVELOPMENT

To develop this project, support tools will be needed to carry out scenario projections, technical drawings, simulation models, and coding of the items listed in Table 1.

Table 1: Types of tools involved in the research.

Tool	Concept	Application in the project
Proteus VSM	Software that allows the creation of circuits, simulation, and layout design for analog and digital applications, including microcontrollers.	It will be used for the development of electronic boards and connections.
Solid Works	Software designed for the creation of three-dimensional virtual mechanical prototypes. It also allows for the simulation of the functioning of parts, assemblies, movements, among others.	In the project, the software will be used to model mechanical parts and devices. From three-dimensional prototypes, 2D drawings will be generated and forwarded to the machining process.

PMBOK Guide	It is a set of project management practices organized by the PMI institute and is considered the basis of project management knowledge by professionals in the field.	It will be used to manage the project, from the initial stage to completion.
Visual Studio Code	Source code editor developed by Microsoft for Windows, Linux, and macOS. It includes support for debugging, built-in Git version control, syntax highlighting, and intelligent code completion.	Development environment.
MySQL Workbench	Client for executing SQL queries, system administration, and database modeling, creation, and maintenance through an integrated environment.	Development environment for queries, data structures, and logical models.
Astah Community	Class, Use Case, Sequence, Communication, State Machine, Activity, Component, Deployment, and Composite Structure Diagram Designer. Used in Software Engineering.	Class project diagrammer and requirements diagrams.
Draw io	Online graphic editor that allows you to develop drawings, graphics, and other designs without the need for expensive and heavy software. It provides resources for creating any type of design and has a section dedicated to information architecture.	Architecture project diagrammer and flowcharts.
Trello	Collaboration tool that organizes projects into boards. At a glance, Trello shows what is being worked on, who is working on what, and where something is in a process.	Activity manager.
Github Desktop	Application that allows you to interact with GitHub using a GUI instead of the command line or a web browser, with which you can upload, extract, and clone remote repositories.	Code versioning.
UX Prototyper - Justinmind	Justinmind is a prototyping tool for websites, software applications, and mobile applications that can work with Windows and Mac, or with iOS and Android. There is a free version that you can work with very well.	Design interface prototyper.
Barcode reader for meter tracking	Industrial scanner for reading 1D and 2D barcodes from 15 cm to 15 m	It will be necessary to track the item by station and identify through data where the product is located, whether it has been rejected or not, in addition to adding the completed stage information to the block body according to the line layout.

Source: Authors, (2026).

III.3 TECHNOLOGIES FOR PROJECT DEVELOPMENT

The technologies that will be used and applied in the project meet the needs of supplying the technology presented as the scope of the project, Internet of Things (IoT), Intelligent Systems, and Artificial Intelligence.

Table 2: Technologies applied in the research process.

Technology	Concept	Application in the project
Vue.js	Vue.js is an open-source JavaScript framework focused on developing user interfaces and single-page applications.	JavaScript framework for building user interfaces.
Vue Router	Vue Router is the official router for Vue.js. It integrates deeply with the core of Vue.js to make it easy to build Single Page Applications with Vue.	Library for page routing.
Vue-118n	Vue I18n is a plugin for internationalization in Vue.js. It easily integrates localization features into your Vue.js application.	Library for internationalizing Vue.js applications.
VueUse	It is a collection of hundreds of essential Vue composition utilities for interacting with various browsers, sensors, animations, and state APIs.	Hooks library for Vue.js.
Quasar or Naive UI	It is an open-source framework based on Vue.js for building applications with a single code base. It can be deployed on the web as SPA, PWA, SSR, in a mobile application using Cordova for iOS and Android, and in a desktop application using Electron for Mac, Windows, and Linux.	Vue.js component framework.
Echart	It is an open-source JavaScript visualization tool that can run fluently on desktops, web apps, and mobile devices.	Data visualization library for creating graphs.
Moment.js	It is a very popular JavaScript library that facilitates working with dates and times in projects developed in this language. With it, you can perform tasks such as manipulating, formatting, and validating dates more simply and efficiently.	Library for manipulating dates and times.
Axios	Axios is a library for making HTTP requests and handling data communication between the frontend and backend in web applications. It provides a simple and flexible API for making asynchronous API calls, such as getting data from a RESTful API, sending data to a server, or updating data in real time.	HTTP client for communicating with the back end.
FastAPI	It is a web framework first released in 2018 for building HTTP-based service APIs in Python. It is used to build APIs with Python 3.8+ based on standard Python type hints. FastAPI is based on Pydantic and uses type hints to validate, serialize, and deserialize data.	Web framework.
SQLAIchemy	It is a cross-platform framework that has quickly become one of the most widely used object-relational mapping tools in the Python community, the language for which it was developed. It provides full SQL flexibility,	ORM library for interacting with databases in Python.

	obtaining a complete set of well-known enterprise-level persistence patterns that are designed for efficient, high-performance database access.	
Redis	It is used by companies that need high-performance tools due to the processing of huge volumes of data, such as Twitter, or applications with real-time updates, such as betting sites.	In-memory database for caching.
Celery	It is a robust, simple, and flexible distributed system that allows you to execute large numbers of messages asynchronously. It is a task queue focused on real-time processing, in addition to supporting task scheduling.	Library for executing asynchronous tasks.
Websockets	WebSocket is a technology that enables bidirectional communication over full-duplex channels on a single Transmission Control Protocol socket. It is designed to run on browsers and web servers that support HTML5, but can be used by any application client or server.	Bidirectional communication protocol over HTTP.
Pydantic	Pydantic Python is a powerful library for Python developers, offering features such as data validation, automatic type conversion, data serialization and deserialization, and automatic documentation.	Library for data validation in Python.
Pytest	Pytest is a Python testing framework that originated from the PyPy project. It can be used to write various types of software tests, including unit tests, integration tests, end-to-end tests, and functional tests. Its features include parameterized tests, fixtures, and assertive rewriting.	Testing framework for Python.

Source: Authors, (2026).

IV. RESULTS AND DISCUSSIONS

The execution of planning, coordination, and technical structuring activities throughout the period resulted in significant advancements in both administrative organization and the consolidation of the technical foundation required for system development. The main outcomes achieved include: validation of the technical scope with the company; alignment of execution and delivery schedules; structuring of the development team after a rigorous selection process; establishment of versioning and collaboration guidelines via GitHub; strategic prioritization of the Technical Sheets module; and practical immersion in the partner company's manufacturing environment, enabling full understanding of production flows and critical process points. From these results, several relevant conclusions were drawn. First, the complexity of the company's industrial processes requires a highly integrated and flexible approach capable of adapting to the demands of traceability, compliance, and automation. Second, replacing manual processes with integrated digital modules represents not only an operational necessity but also a concrete opportunity for gains in efficiency, standardization, and reliability. It was also found that the integration among the research's technical teams was essential for advancing technical definitions, demonstrating that multidisciplinary collaboration is a key factor for the project's success.

Among the deliverables produced in this activity, the following stand out:

• Technical diagram and mind map of the functional structure of the Technical Sheets module;
• Functional mockups of the SIPE system interfaces;
• Restructured schedule based on the new team composition;
• Individual and consolidated technical reports with documentary and visual evidence from technical visits;
• Versioned repositories with code documentation on GitHub;
• Initial architectural definitions involving the use of technologies such as FastAPI, Kafka, MariaDB, and ClickHouse;
• Technical visit reports mapping industrial processes and integration points with the integrated system.

In terms of investigation and experimentation, the activity validated the technical feasibility of digitizing and automating the Technical Sheets generation process, based on a critical review of the documents currently used by the company. The analysis of production flows—especially in the meter assembly, plastic injection, and metrological laboratory sectors—provided practical input for refining system requirements. Industrial communication technologies such as MQTT, HTTP, RS-232, RS-485, and IEC 62056-21 optical interfaces were also examined, confirming the feasibility of integration between physical devices and management systems. The activity significantly contributed to increasing the knowledge of the professionals involved regarding Industry 4.0 enabling technologies, particularly in industrial process automation, digital traceability, system integration, real-time data management, and information security. Coordinators and developers participated in practical training and immersion activities within the factory processes, gaining applied knowledge about digitizing technical documents, API modeling, structuring hybrid databases (OLTP and OLAP), scalable backend architecture, and interoperability standards with industrial ERPs. This continuous learning strengthened the team's technical capabilities and raised the project's overall technological maturity, aligning it with the principles of Industry 4.0.

The results obtained in the seventh month can be analyzed across three dimensions: qualitative, quantitative, and organizational. Each provides important insights into understanding the impact of the activities carried out and the added value brought to the project. In the qualitative dimension, progress was significant. User experience improved substantially with the redesign of the sidebar and the implementation of the sheet preview feature. The system became more intuitive, reducing learning curves and increasing operational efficiency. Additionally, standardization through templates brought greater reliability to the recorded information, reducing human error and strengthening traceability within production processes From a quantitative perspective, the numbers also demonstrate progress. More than twenty parameterized templates were created during the month, forming a solid foundation for different application scenarios. The average time required to complete the sheets was reduced by approximately thirty percent, showing the direct impact of automation and standardization on the team's daily routine. Furthermore, fifteen critical inconsistencies identified in previous cycles were corrected, reinforcing the maturity of the validation process and improving system reliability, as shown in Figure 4.

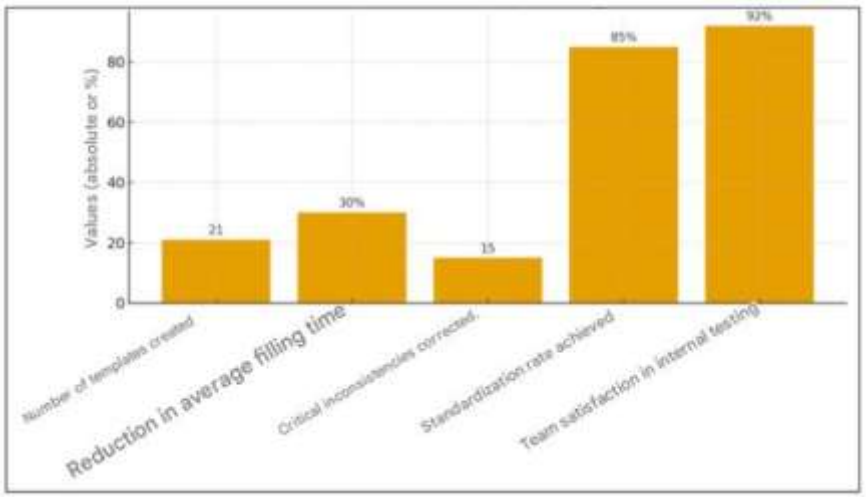


Figure 4: Quantitative performance indicators for the Technical Data Sheet module in the 7th month.
Source: Authors, (2026).

Following the implementation of product traceability, which was the focus of phase 1 of development, this new phase aims to elevate the system to a more comprehensive MES level, with greater control over production flow and production order management. This includes real-time monitoring of the production plan, allowing the system to actively interact with the equipment, determining what should be produced, reducing dependence on production operators, improving process quality, and applying stricter control over the production plan and quantities produced. To interact with production systems and equipment, various documents will be transferred to the SIPE system to avoid ambiguity and errors in data entry, such as labels, laser markings, and proper product configuration according to customer specifications. Figure 5 shows the hierarchical levels of industrial systems, organized in a pyramid ranging from machines (physical level) to ERP (corporate level). It shows the expected reaction times at each layer and the types of events managed, such as failures and rejections. This structure facilitates integration between the shop floor and management systems.

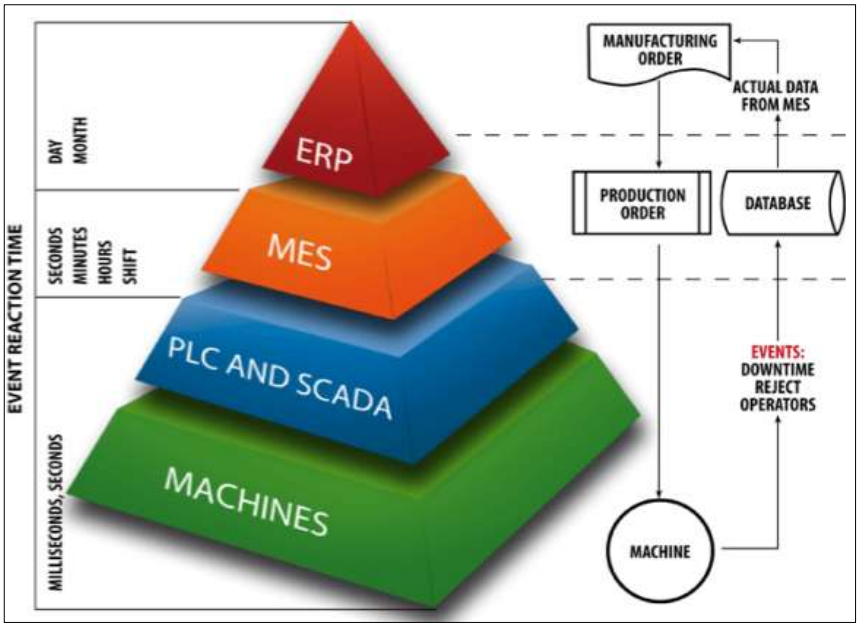


Figure 5: Layered Architecture of Industrial Automation.
Source: Authors, (2026).

IV.1 TECHNOLOGICAL INNOVATIONS

The SIPE System features process innovation, being a new solution for the company, for the Western Amazon region, and for Brazil. The main innovations include:

a. IoT-based traceability devices;
b. Intelligent production planning system with real-time data interpretation;
c. Artificial intelligence algorithms for dynamic production replanning;
d. Dashboards integrated with Business Intelligence (BI);
e. Integration via microservices and APIs for modularity and scalability;
f. Predictive production control based on industrial data analysis.

The system uses a simple yet powerful design approach, utilizing modern development SDKs for easy maintenance and improvement. All system functions and transactions are performed on the server side, using a REST API to interact with equipment for data collection and control. The following figure shows the SIPE system architecture diagram using Vue.js on the front end, Django (Python) on the back end, MariaDB database, Redis cache, and containerization with Docker. Communication between client and server occurs via REST API using data in JSON format, ensuring scalability, modularity, and performance in industrial applications. In Figure 6, the diagram is presented that connects template versions to an active workflow with sequential levels, where each level defines the minimum number of approvers and who is allowed to approve by role or by individual. Decisions are recorded per level, with a full audit trail. The setup sheet inherits the dynamic structure and directly references the original technical sheet, preserving traceability between the approved specification and the configuration executed on the production line. In this stage, a set of unit and integration tests was also conducted to validate the robustness of the Technical Sheet approval module and the derivation of the Setup Sheet. The focus was on ensuring logical consistency, data integrity, and performance under load, covering the following main scenarios:

• **Workflow:** Execution of scenarios with multiple approval levels, including approvers defined by role or by specific user. Chain rejections and alternative flows were simulated to validate system behavior in cases of partial quorum or lack of consensus.
• **Validations:** Tests for duplicate approvals by the same agent, status transitions (pending → approved/rejected), and automatic blocks when the minimum number of approvers was not reached. Business rules were validated with audit records for each event.
• **Setup Sheet:** Verification of correct data inheritance from the approved Technical Sheet, tests for submission of configuration parameters, and checklist registration. Traceability between the original Technical Sheet and the executed Setup was audited.
• **Integration and Performance:** REST endpoints related to the workflow and the Setup Sheet were subjected to moderate-load scenarios, evaluating response consistency, average latency, and stability under multiple concurrent requests.

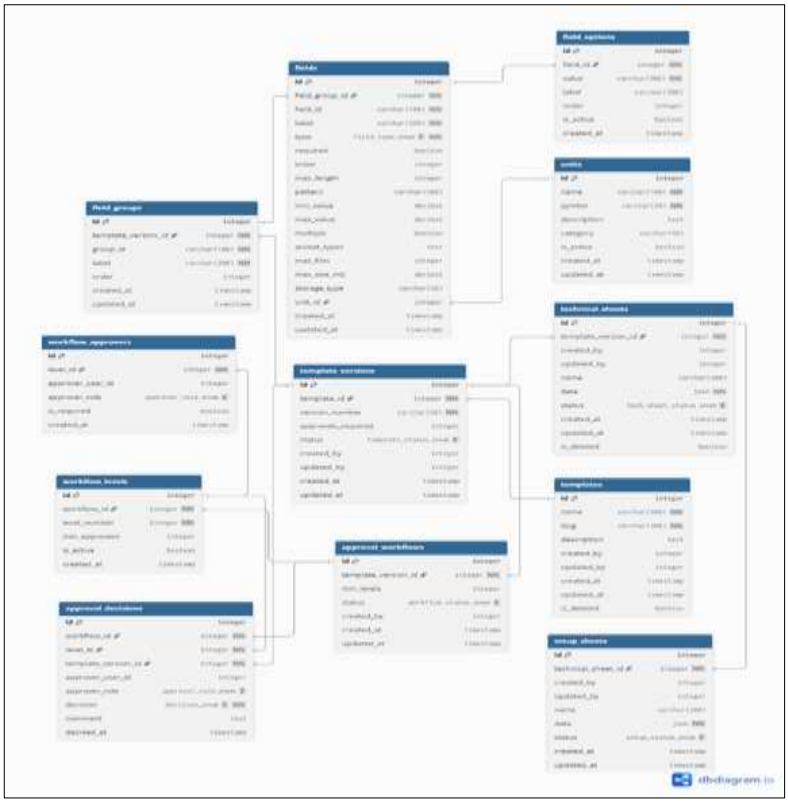


Figure 6: Multi-level approval workflow and derived setup form.
Source: Authors, (2026).

Table 3 below summarizes the main deliverables and improvements achieved in this development cycle.

Table 3: Improvement and delivery of the main activities of this sub-activity.

Delivery	Description	Benefit
Approval workflow	Multi-level flow with dynamic approvers	Document governance and traceability
Validation rules	Minimum number of approvers, specific positions/users	Reliability and security in the decision-making process
Auditing	Record of identity, position, decision, and date/time	Transparency and complete history
Setup sheet	Derived directly from the approved Technical Data Sheet	Traceable integration with machine configuration
Setup fields	Parameters, calibration, and operational checklists	Compliance and standardization in line preparation
REST integration	Endpoints for workflow and Setup Sheet	Standardized operability and secure consumption by external systems
Load testing	Tests with multiple approvers and stress scenarios	Stable and predictable performance
Duplicate prevention	Blocking of repeated approvals by the same agent	Integrity and fairness in the approval process

Source: Authors, (2026).

During the development cycle, critical points were identified and addressed, as they directly affected the system's consistency. Among the detected errors, the most notable was approval concurrency within the same level, which caused status inconsistencies when multiple approvers submitted decisions simultaneously. This issue was resolved through the adoption of transactional control with locks, ensuring the atomicity of operations. Another recurring error was the failure in linking the Setup Sheet to the Technical Sheet in cases of partial submission, which was corrected by enforcing mandatory referential integrity constraints, preventing setups from being created without a valid reference to the originating technical document. Duplicate approvals by the same user at different levels were also observed; this issue was corrected by implementing uniqueness validation per level, complemented by more robust audit trails. On the other hand, the successful aspects reinforced the solidity of the architecture. The workflow demonstrated high flexibility by allowing parameterization by roles and specific users, while maintaining adherence to business rules. The requirement for mandatory derivation of the Setup Sheet from an approved Technical Sheet ensured traceability between specification and factory execution. The APIs were standardized and submitted to integration tests, ensuring interoperability and consistency in external consumption. Furthermore, the complete audit trails—recording identity, role, decision, and timestamp—strengthened document governance. Finally, the endpoints maintained stable performance even under load scenarios, validating the system's scalability.

IV.2 HARDWARE AND BACKEND INTEGRATION

With the arrival and validation of the Raspberry Pi 5 and ESP32 hardware, the transition began from the code previously tested in simulation to the factory environment. This stage was carried out using an incremental integration approach, including bench tests and fine adjustments to communication protocols to ensure stability and adherence to real operating conditions. The first focus was on the ESP32 communication module, which was configured with firmware capable of supporting multiple communication channels, performing UART tests for interaction with sensors and actuators, as well as direct synchronization with the Raspberry Pi 5. The priority was to achieve low latency and deterministic response. To this end, a bottom-up methodology was adopted, in which lower-level routines were validated individually before assembling the main flow. Next, the Raspberry Pi 5 was integrated with the factory backend, receiving routines for collecting processed data from the ESP32 and sending it via REST API, with packet compression and exception handling. Dynamic authentication tokens were implemented in accordance with security guidelines, and load scenarios with multiple simultaneous connections were evaluated, simulating high-demand factory conditions. The integration followed Continuous Integration (CI) practices, with incremental builds and automated tests for each modification to drivers or communication routines.

Subsequently, synchronization tests were performed between the hardware and the factory system, validating consistency between the Raspberry Pi's local records and the data stored in the Horus database. Watchdogs were also implemented on the ESP32 to enable automatic reboot in case of communication failures, and peak-production tests were conducted with multiple meters connected. This phase followed principles of fault tolerance and minimal redundancy, supported by detailed logs for auditing and performance analysis. The results confirmed the robustness of the architecture, real-time traceability in Horus, and resilience against communication failures, preparing the solution for the next stage: the Production Control Module, focused on continuous data collection and metrics calculation such as productivity, downtime, losses, and rework. In parallel with the hardware integration, refinement of the template management module was carried out, focusing on code modularization, removal of redundancies, and performance optimization, ensuring greater structural clarity and maintainability. In terms of UX/UI, adjustments were applied to improve visual consistency and simplify navigation flows, reinforcing adherence to the standards established for the system.



Figure 7: Folders with refactored modules.
Source: Authors, (2026).

Critical bugs identified in template creation were corrected, including field validation failures, integration issues, and inconsistencies found in regression tests. The system's visual standardization was revised, reinforcing responsiveness for different devices. In this context, the Sidebar was redesigned with a scalable layout, optimized hierarchy, and more intuitive navigation, strengthening the visual identity and coherence of the frontend.

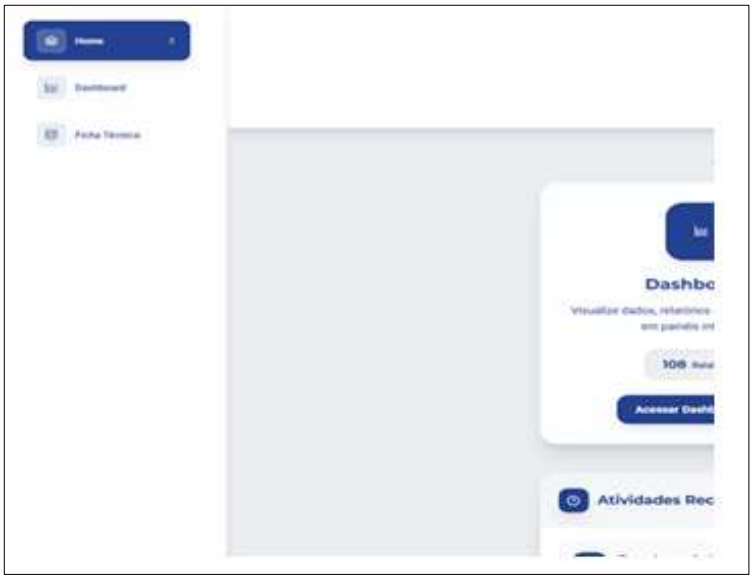


Figure 8: Sidebar.
Source: Authors, (2026).

The integration of the new functionalities into the backend was validated with revised API contracts and end-to-end tests. The DBeaver tool was extensively used to inspect data persistence, verify referential integrity, and confirm alignment between the user interface and the records actually stored in the database.

V. CONCLUSIONS

The execution of the project demonstrated significant progress in consolidating the Technical Data Sheets module, evidenced by the creation of a detailed technical diagram and a mind map that clearly represent the system's functional structure. These tools enabled the visualization of relationships between components and the flow of information, facilitating both the understanding of the module and the identification of potential improvements. In parallel, the functional mockups of the developed system's interfaces provided a solid foundation for validating the user experience, ensuring that the planned functionalities are aligned with operational needs and the manufacturing processes observed during technical visits. The research schedule was restructured according to the new team composition, ensuring optimized resource allocation and the achievement of established deadlines. Technical reports, both individual and consolidated, systematically documented the technical visits, presenting visual and descriptive evidence of the practices and mapped processes, which allowed the identification of critical integration points for the integrated system.

Furthermore, the maintenance of versioned repositories on GitHub ensured efficient control of the source code and associated documentation, promoting transparency and ease of tracking the development process. The initial architectural definitions, involving technologies such as FastAPI, Kafka, MariaDB, and ClickHouse, provided a robust foundation for scalable and efficient system development, ensuring interoperability and reliability in data processing. The detailed mapping of manufacturing processes and the identification of integration points with SIPE reinforce the importance of a structured, process-oriented approach, allowing the Technical Data Sheets module to effectively align with existing operations and support decision-making based on accurate and up-to-date information. In summary, the results obtained so far demonstrate consistent progress, technological integration, and adherence to best development practices, establishing a solid foundation for the next phases from the survey.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Jandecy Cabral Leite, Marcelo Maia do Nascimento, Alan Ferreira Pinheiro Tavares, André Barbosa de Lima, Matheus Pereira de Souza, Daniel Gomes Gulart and Marcos Castro de Menezes.
Methodology: Alan Ferreira Pinheiro Tavares, André Barbosa de Lima, Matheus Pereira de Souza.
Investigation: Jandecy Cabral Leite, Marcelo Maia do Nascimento, Alan Ferreira Pinheiro Tavares, André Barbosa de Lima, Matheus Pereira de Souza, Daniel Gomes Gulart and Marcos Castro de Menezes.
Discussion of results: Jandecy Cabral Leite, Marcelo Maia do Nascimento, Alan Ferreira Pinheiro Tavares.
Writing – Jandecy Cabral Leite.
Writing – Review and Editing: Jandecy Cabral Leite and Marcelo Maia do Nascimento.
Resources: Jandecy Cabral Leite, Marcelo Maia do Nascimento.
Supervision: Jandecy Cabral Leite, Marcelo Maia do Nascimento and Alan Ferreira Pinheiro Tavares.
Approval of the final text: Jandecy Cabral Leite, Marcelo Maia do Nascimento.

VII. ACKNOWLEDGEMENTS

To ITEGAM and Wasion da Amazônia Electronic Instruments Industry LTDA for the financial support of the R&D&I Project through the Information Technology Law (Law No. 8,248/1991), which encourages investment in research, development, and innovation (R&D&I) in the information and communication technology (ICT) sector through tax incentives and financial credits for companies in the Manaus Industrial Pole (PIM).

VIII. REFERENCES

[1] Borangiu, Theodor et al. Digital transformation of manufacturing. Industry of the future with cyber-physical production systems. Romanian journal of information science and technology, v. 23, n. 1, p. 3-37, 2020.

[2] Tsaramirsis, Georgios et al. A modern approach towards an industry 4.0 model: From driving technologies to management. Journal of Sensors, v. 2022, n. 1, p. 5023011, 2022.

[3] Yaqub, Muhammad Zafar; Alsabban, Abdullah. Industry-4.0-enabled digital transformation: Prospects, instruments, challenges, and implications for business strategies. Sustainability, v. 15, n. 11, p. 8553, 2023.

[4] Xu, Hansong et al. A survey on industrial Internet of Things: A cyber-physical systems perspective. Ieee access, v. 6, p. 78238-78259, 2018.

[5] Radanliev, Petar et al. Artificial intelligence in cyber physical systems. AI & society, v. 36, n. 3, p. 783-796, 2021.

[6] Tran-Dang, Hoa; Kim, Dong-Seong. The physical internet in the era of digital transformation: perspectives and open issues. Ieee Access, v. 9, p. 164613-164631, 2021.

[7] Salau, Babajide A.; Rawal, Atul; Rawat, Danda B. Recent advances in artificial intelligence for wireless internet of things and cyber–physical systems: A comprehensive survey. IEEE Internet of Things Journal, v. 9, n. 15, p. 12916-12930, 2022.

[8] Chander, Bhanu et al. Artificial intelligence-based internet of things for industry 5.0. In: Artificial intelligence-based internet of things systems. Cham: Springer International Publishing, 2022. p. 3-45.

[9] Zhang, Kunwu et al. Advancements in industrial cyber-physical systems: An overview and perspectives. IEEE Transactions on Industrial Informatics, v. 19, n. 1, p. 716-729, 2022.

[10] Taj, Imran; Zaman, Noor. Towards industrial revolution 5.0 and explainable artificial intelligence: Challenges and opportunities. International Journal of Computing and Digital Systems, v. 12, n. 1, p. 295-320, 2022.

[11] Bueno, Adauto; Godinho Filho, Moacir; Frank, Alejandro G. Smart production planning and control in the Industry 4.0 context: A systematic literature review. Computers & industrial engineering, v. 149, p. 106774, 2020.

[12] Luo, Dan; Thevenin, Simon; Dolgui, Alexandre. A state-of-the-art on production planning in Industry 4.0. International Journal of Production Research, v. 61, n. 19, p. 6602-6632, 2023.

[13] Bendul, Julia C.; Blunck, Henning. The design space of production planning and control for industry 4.0. Computers in Industry, v. 105, p. 260-272, 2019.

[14] Usuga Cadavid, Juan Pablo et al. Machine learning applied in production planning and control: a state-of-the-art in the era of industry 4.0. Journal of Intelligent Manufacturing, v. 31, n. 6, p. 1531-1558, 2020.

[15] Herrmann, Jan-Phillip et al. Approaches of production planning and control under industry 4.0: A literature review. Journal of Industrial Engineering and Management (JIEM), v. 15, n. 1, p. 4-30, 2022.

[16] Gao, Qiang; Li, Fei; Chen, Chong. Research of internet of things applied to manufacturing execution system. in: 2015 ieee international conference on cyber technology in automation, control, and intelligent systems (cyber). ieee, 2015. p. 661-665.

[17] De Lima Souza, Karina et al. Desenvolvimento de um sistema de execução de manufatura (MES) no planejamento e controle de produção: uma aplicação da Indústria 4.0 no processo de fabricação de tubos de aço. Produto & Produção, v. 22, n. 1, 2021.

[18] Dutta, Gautam et al. Overcoming the barriers of effective implementation of manufacturing execution system in pursuit of smart manufacturing in SMEs. Procedia Computer Science, v. 200, p. 820-832, 2022.

[19] Alexakos, Christos et al. Iot integration for adaptive manufacturing. In: 2018 IEEE 21st International Symposium on Real-Time Distributed Computing (ISORC). IEEE, 2018. p. 146-151.

[20] Kim, Tae Hyun; Jeong, Jongpil; Kim, Yeasang. A conceptual model of smart manufacturing execution system for rolling stock manufacturer. Procedia Computer Science, v. 151, p. 600-606, 2019.

[21] Almada-Lobo, Francisco. The Industry 4.0 revolution and the future of Manufacturing Execution Systems (MES). Journal of innovation management, v. 3, n. 4, p. 16-21, 2015.

[22] Jaskó, Szilárd et al. Development of manufacturing execution systems in accordance with Industry 4.0 requirements: A review of standard-and ontology-based methodologies and tools. Computers in industry, v. 123, p. 103300, 2020.

[23] Barari, Ahmad et al. Intelligent manufacturing systems towards industry 4.0 era. Journal of Intelligent Manufacturing, v. 32, n. 7, p. 1793-1796, 2021.

[24] Ramesh, R.; Jyothirmai, S.; Lavanya, K. Intelligent automation of design and manufacturing in machine tools using an open architecture motion controller. Journal of Manufacturing Systems, v. 32, n. 1, p. 248-259, 2013.

[25] Coronado, Pedro Daniel Urbina et al. Part data integration in the Shop Floor Digital Twin: Mobile and cloud technologies to enable a manufacturing execution system. Journal of manufacturing systems, v. 48, p. 25-33, 2018.

[26] Mantravadi, Soujanya; Møller, Charles. An overview of next-generation manufacturing execution systems: How important is MES for industry 4.0?. Procedia manufacturing, v. 30, p. 588-595, 2019.

[27] Freakley, Philip K. Rubber processing and production organization. Springer Science & Business Media, 2012.

[28] Knafl, Matthew et al. Diagnosing PCP failure characteristics using exception based surveillance in CSG. In: SPE progressing cavity pumps conference. SPE, 2013. p. SPE-165655-MS.

[29] Gonzalez, L. E.; Chokshi, R. N.; Lane, W. C. Real-Time Surface and Downhole Measurements and Analysis for Optimizing Production. In: SPE Asia Pacific Oil and Gas Conference and Exhibition. SPE, 2015. p. SPE-176233-MS.

[30] Delfim, Leonardo Melo et al. Measuring costs in a containerized cargo export chain: analysis from the perspective of an exporting company. Gestão & Produção, v. 30, p. e12422, 2023.

[31] Özengin, Fatih. Accounting Information Systems and Re-Engineering of Inventory System Within Mrpii (Manufacturing Resource Planning) of an Industrial Company. 1997. Dissertação de Mestrado. Marmara Universitesi (Turkey).

[32] Oceania, Siti Annisafa et al. Procurement of Ship Materials Through the Shipmate Application using the Material Requirement Planning Method. 2024.

[33] Richard, Thomson; Gozali, Lina; Daywin, Frans Jusuf. Management of raw material needs and safety stock based on data forecast and system dynamics modeling. In: 2023 19th IEEE International Colloquium on Signal Processing & Its Applications (CSPA). IEEE, 2023. p. 7-12.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 496-518. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3146

RESEARCH ARTICLE OPEN ACCESS

AUTOMATED CONTROL AND PLANNING SYSTEM FOR THE PRODUCTION OF ELECTRONIC METERS WITH ARTIFICIAL INTELLIGENCE

Jandecy Cabral Leite[1], Marcelo Maia do Nascimento[2], Marivan Silva Gomes[3], Alan Ferreira Pinheiro Tavares[4], André Barbosa de Lima[5], Matheus Pereira de Souza[6], Daniel Gomes Gulart[7] and Marcos Castro de Menezes[8]

[1, 2,3,4,5,6]Department of Industrial Research Department of the Galileo Institute of Technology and Education of Amazonia (ITEGAM). Address: Joaquim Nabuco, No. 1950. Neighborhood: Downtown. ZIP CODE: 69020-030. Manaus-Amazonas, Brazil.
[7,8]Wasion da Amazônia Electronic Instruments Industry LTDA. Av. Cosme Ferreira, No. 12.110. Colônia Antônio Aleixo. ZIP CODE: 69008-310. Manaus – Amazonas, Brazil.

[1]https://orcid.org/0000-0002-1337-3549, [2]http://orcid.org/0000-0001-7183-8642, [3]http://orcid.org/0009-0007-2022-9047
[4]http://orcid.org/ 0000-0003-1644-9999, [5]http://orcid.org/0009-0002-6910-0635, [6]http://orcid.org/0009-0008-3447-3566
[7]http://orcid.org/0009-0002-2787-7753, [8]http://orcid.org/0009-0000-7874-9132

Email: *jandecy.cabral@itegam.org.br, marcelo.nascimento@itegam.org.br, msgomes@uea.edu.br, alan.tavares@itegam.org.br, amilandre@gmail.com, matheus@itegam.org.br, daniel.gulart@wasion.com.br, marcos.castro.menezes@gmail.com

ARTICLE INFO

Article History
Received: December 15, 2025
Revised: December 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Industry 4.0, Artificial Intelligence (AI), Production Planning and Control, Electronic Meters, Operational Optimization, Automated System.

ABSTRACT

In the context of Industry 4.0, modern manufacturing environments increasingly rely on digital technologies to enhance efficiency and competitiveness. The production of electronic meters represents a complex industrial process that demands precise coordination of resources, materials, and scheduling. Traditional methods of production planning often face challenges such as bottlenecks, resource underutilization, and delayed delivery times. To address these issues, the integration of Artificial Intelligence (AI) into manufacturing processes has emerged as a promising solution. The main objective of this research is to develop an automated system for production planning of electronic meters that leverages AI techniques to optimize operational performance. The study proposes a model capable of analyzing historical production data, predicting demand fluctuations, and dynamically adjusting production schedules. The methodology involves the implementation of machine learning algorithms and optimization techniques within a simulated industrial environment. The system evaluates different production scenarios, prioritizes tasks, and allocates resources efficiently to meet production targets while minimizing delays. Results demonstrate that the proposed AI-based automated system significantly improves operational optimization. Key performance indicators, such as production throughput, resource utilization, and lead time, showed measurable enhancements compared to conventional planning methods. Moreover, the system enables real-time decision-making, which aligns with the principles of Industry 4.0 and supports adaptive, data-driven production strategies. In conclusion, the research highlights the potential of combining Artificial Intelligence and automated systems to advance production planning for electronic meters, contributing to more efficient, flexible, and responsive manufacturing processes in the Industry 4.0 era.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The rapid technological evolution of the industrial sector has accelerated the transition toward Industry 4.0, a paradigm characterized by digital integration, cyber-physical systems, data analytics, and intelligent automation.

In this new industrial landscape, manufacturing processes increasingly depend on advanced computational tools capable of enhancing efficiency, adaptability, and connectivity. The production of electronic meters, in particular, has faced growing challenges due

to market expansion, demand variability, and the need for continuous improvement in quality and delivery performance. Traditional methods of Production Planning and Control often prove insufficient to handle the complexity of modern manufacturing environments, leading to inefficiencies such as delays, resource imbalances, and limited responsiveness to unforeseen changes [1-5]. To address these limitations, the introduction of Artificial Intelligence (AI) has emerged as one of the most transformative innovations in industrial operations. AI techniques—especially machine learning, optimization algorithms, and predictive analytics—offer the ability to process large volumes of data, identify patterns, and support decision-making with greater accuracy and speed. When applied to Production Planning and Control, AI enables systems to dynamically adjust schedules, predict bottlenecks, and optimize resource allocation in real time.

This creates a foundation for intelligent, flexible, and highly efficient manufacturing systems. In this context, the development of an automated system designed to enhance production strategies for electronic meters becomes essential. By integrating AI algorithms with industrial data streams, such a system can significantly contribute to operational optimization, reducing lead times, increasing production throughput, and minimizing waste. Furthermore, automation allows organizations to adapt more effectively to fluctuations in demand and operational disturbances, strengthening competitiveness in a globalized market [6-9]. Overall, the incorporation of Industry 4.0 concepts—combined with Artificial Intelligence and automated decision-making—marks a critical step toward transforming production environments. The application of these technologies to the manufacturing of electronic meters not only supports smarter Production Planning and Control but also aligns industrial processes with the current demands for agility, precision, and sustainability. This introduction lays the groundwork for exploring how intelligent systems can enhance industrial performance and redefine the future of automated production.

II. LITERATURE REVIEW

II.1 INDUSTRY 4.0

Industry 4.0 represents the fourth industrial revolution and is characterized by the deep integration of digital technologies, physical systems, and intelligent production processes. It emerged in response to the growing need for efficiency, customization, speed, and flexibility in modern manufacturing chains. Unlike previous revolutions, Industry 4.0 promotes the convergence of advanced automation, data analytics, interconnectivity, and computational intelligence, transforming traditional factories into fully connected and autonomous environments [10],[11]. Among its main characteristics is the adoption of cyber-physical systems capable of monitoring and controlling industrial processes in real time. These systems communicate with each other through the Internet of Things (IoT), allowing machines, sensors, and software to continuously exchange information. This connectivity creates a more transparent, efficient, and predictable production network. In addition, the use of Big Data and advanced analytics makes it possible to interpret large volumes of data generated during production, supporting strategic and operational decisions with greater accuracy. Within this ecosystem, Artificial Intelligence (AI) plays a central role. AI enables machines to learn patterns, predict failures, optimize routines, and make autonomous decisions. In industrial processes, machine learning algorithms assist in production planning, predictive maintenance, automated quality inspection, and performance analysis. AI also helps dynamically adjust production lines, reducing waste, increasing productivity, and ensuring greater reliability of results [12-15]. Another important characteristic of Industry 4.0 is horizontal and vertical integration. Horizontal integration connects different companies along the production chain, while vertical integration connects all levels of the organization—from the shop floor to corporate management. This integration ensures greater synchronization and efficiency. Thus, Industry 4.0 represents a transformative milestone, driving smart factories, data-driven decision-making, and a more flexible, sustainable, and competitive production model.

II.2 ARTIFICIAL INTELLIGENCE (AI)

Artificial Intelligence (AI) has established itself as a central element in modern technological advancement, especially when combined with other tools that make up the digital transformation ecosystem. Among these tools, the Internet of Things (IoT), Big Data, Machine Learning, and several techniques derived from advanced data analytics stand out. IoT enables continuous connectivity between machines, sensors, devices, and systems, allowing real-time data collection directly from the physical environment. This constant monitoring capability generates an enormous amount of information, which is subsequently stored and processed using Big Data technologies. Processing these large volumes of data is essential for AI to extract useful knowledge, identify patterns, and support faster and more accurate decision-making [16-18]. Within this context, Machine Learning—one of the main branches of Artificial Intelligence—plays a crucial role. Through learning algorithms, systems are able to analyze historical data, recognize trends, and continuously learn without direct human intervention. These techniques are widely used in industrial applications such as demand forecasting, predictive maintenance, fault detection, and process optimization. In addition, other complementary techniques, such as Deep Learning, Process Mining, Predictive Analytics, Expert Systems, and Cloud Computing, further expand the potential of AI applications. Deep Learning, for example, enables highly complex analyses and is used in computer vision and pattern recognition. Cloud computing ensures scalability, efficient storage, and processing for AI-based systems. The integration of all these technologies creates an intelligent and autonomous environment capable of improving operations, reducing costs, and increasing efficiency across various sectors, especially in modern industry. This synergy has driven the transition toward more connected, predictable, and data-driven production systems [19-21].

II.3 PRODUCTION PLANNING AND CONTROL

Production Planning and Control (PPC) is a set of essential practices to ensure that the production process takes place in an organized, efficient manner and is aligned with market needs. In industrial environments, PPC seeks to integrate resources, deadlines, productive capacities, and demand, ensuring that production occurs at the right time, in the desired quantity, and at the lowest possible cost. The first phase is planning, which involves demand forecasting, defining production targets, and estimating the required resources. At this stage, sales history, seasonality, inventory levels, and machine capacities are analyzed. Based on this information, the Master Production Schedule (MPS) is developed, a document that guides what to produce, when to produce it, and in what quantity.

The second phase is materials planning, also known as MRP (Material Requirements Planning). It determines which materials will be needed, in what quantities, and when they must be available. The objective is to avoid shortages that interrupt production and excess inventory that generates storage costs. Next comes capacity planning, which assesses whether productive resources—machines, teams, and available time—are sufficient to meet the plan. If constraints are identified, adjustments to the schedule or redistribution of the workload are made. The production scheduling phase defines the sequence of operations on the manufacturing line, optimizing machine usage and minimizing bottlenecks. This stage aims to establish clear and efficient production orders [22-25]. Finally, production control monitors execution in real time, comparing planned results with actual performance. Indicators such as productivity, efficiency, cycle times, and delays are analyzed to support quick decision-making. Thus, Production Planning and Control plays a fundamental role in industrial competitiveness, ensuring more stable, predictable, and economically efficient production processes.

II.4 ELECTRONIC METERS

Electronic Meters are essential devices used to measure, monitor, and record the consumption of electrical energy in residential, commercial, and industrial environments. Unlike traditional electromechanical meters, electronic meters operate using digital components and advanced sensing technologies capable of offering greater precision, real-time monitoring, and improved communication capabilities. Their evolution is strongly linked to the modernization of electrical grids and the global movement toward smarter, more automated energy management systems. One of the key characteristics of electronic meters is their ability to collect high-resolution data. Through embedded microprocessors and electronic sensors, these devices measure voltage, current, power factor, instantaneous consumption, and other electrical parameters with superior accuracy. This precise data collection supports utilities in analyzing consumption patterns, identifying losses, and improving demand forecasting. For consumers, it enables a clearer understanding of their energy usage, encouraging more conscious and efficient consumption habits [26-28]. Modern electronic meters also integrate communication technologies that allow remote data transmission. Through protocols such as RF mesh, PLC (Power Line Communication), or cellular networks, these meters can send consumption information directly to utility companies without the need for manual readings.

This feature forms the basis of Advanced Metering Infrastructure (AMI), which supports bidirectional communication and enables functionalities such as remote connection and disconnection, tariff updates, and automatic reporting of anomalies. In addition to basic measurement functions, electronic meters can include advanced features such as load profiling, event logging, tamper detection, and power quality monitoring. These capabilities increase the reliability of the electrical system by identifying irregularities like voltage dips, spikes, or unauthorized usage attempts. Electronic meters also play a fundamental role in the digital transformation of power distribution networks, often referred to as smart grids. Their integration with data analytics platforms and management systems enables utilities to optimize grid operations, reduce operational costs, and improve energy distribution efficiency [29-31]. Overall, electronic meters represent a significant advancement in metering technology, supporting intelligent energy management, strengthening grid stability, and contributing to the modernization of the electrical sector. As digital technologies continue to evolve, these devices are expected to become even more interconnected, accurate, and capable of supporting future smart energy solutions.

II.5 OPERATIONAL OPTIMIZATION

Operational Optimization is a central concept in modern management, focused on the continuous improvement of production processes, waste reduction, and increased efficiency in industrial and business environments. With the advance of digital transformation, intelligent systems have become indispensable tools for achieving higher levels of operational performance. These systems use Artificial Intelligence (AI), machine learning, real-time data processing, and automation to analyze operations, predict behaviors, and support strategic decision-making. One of the main differentiators of intelligent systems is their ability to collect and process large volumes of data from sensors, machines, corporate systems, and connected platforms. From this set of information, advanced algorithms can identify hidden patterns, diagnose problems, predict failures, and accurately suggest corrective actions. This allows companies to anticipate risks, reduce maintenance costs, and avoid unexpected downtime, directly contributing to increased productivity. In addition, intelligent systems are capable of dynamically optimizing processes. In production lines, for example, they can automatically adjust parameters such as operating speed, task sequencing, resource allocation, and energy consumption, promoting more efficient use of industrial infrastructure.

This real-time adaptability makes the process more agile, reliable, and aligned with current operating conditions [32-34]. Another fundamental aspect of operational optimization is data-driven decision-making. With analytical dashboards, automated reports, and AI-generated forecasts, managers can quickly evaluate operational performance, identify bottlenecks, and define strategic priorities in an informed manner. This reduces uncertainty and significantly improves the quality of decisions. Intelligent systems also contribute to integration between departments, enabling production, logistics, maintenance, and quality to operate in a synchronized way. Communication between machines and corporate systems enables more efficient workflows, full traceability, and greater control over results. In summary, operational optimization supported by intelligent systems transforms the way companies operate, making them more competitive, productive, and prepared for the challenges of the digital era. These systems not only automate tasks but also expand operational intelligence, fostering innovation and sustainability in industrial processes.

II.6 AUTOMATED SYSTEM

An Automated System plays a fundamental role in modern manufacturing environments, especially in the production of electronic devices, where precision, speed, and reliability are essential. As global demand for electronic products continues to rise, industries face the challenge of maintaining high productivity while ensuring strict quality standards. To meet these requirements, automated systems have been widely adopted, transforming traditional production lines into intelligent, efficient, and highly controlled environments. In the manufacturing processes of electronic products—such as circuit boards, sensors, smart modules, and electronic meters—automation enables repetitive activities to be performed with micrometric precision, reducing human error and ensuring consistency.

Stages such as machine soldering, automatic optical inspection (AOI), surface-mount technology (SMT) assembly, and functional testing are now largely carried out by automated systems capable of operating 24 hours a day. In addition to increasing productivity, Automated Systems contribute significantly to improved quality control. Embedded sensors, high-resolution cameras, and inspection algorithms detect defects in real time, allowing immediate corrections and reducing the number of rejected products. This continuous monitoring capability ensures that electronic components meet the safety and performance specifications required by the market. Another important benefit is the integration of these systems with digital platforms and corporate databases. This connection facilitates the collection and analysis of large volumes of operational data, which are used to optimize workflows, predict failures, and schedule preventive maintenance. Data-driven analysis also allows production parameters to be adjusted automatically, ensuring greater energy efficiency and cost reduction. Furthermore, automated systems promote greater flexibility in production lines. Instead of lengthy manual reconfigurations, robots and intelligent machines can be quickly reprogrammed to accommodate new product models or changes in demand, increasing the agility of the production process [35-38]. In summary, the implementation of an Automated System in electronic manufacturing environments represents a significant step toward intelligent production. These systems enhance operational efficiency, improve final product quality, and prepare industries to compete in an increasingly dynamic and technology-driven market.

III. MATERIALS AND METHODS

The proposed computational methodology is based on the principles of Industry 4.0, incorporating advanced digital technologies, real-time connectivity, and intelligent data analysis to enhance Production Planning and Control (PPC) in the manufacturing of Electronic Meters. The objective is to develop an Automated System capable of optimizing industrial operations through Artificial Intelligence (AI) techniques, ensuring greater efficiency, accuracy, and adaptability in production processes. The first stage of the methodology consists of data acquisition and integration, which is typical of Industry 4.0 environments. Sensors, programmable logic controllers (PLCs), Manufacturing Execution Systems (MES), and IoT platforms continuously collect information on machine performance, material consumption, cycle times, inventory, and production orders. These data are transmitted to a cloud infrastructure or an industrial server, where they are stored and organized. In the second stage, data processing is carried out using Big Data tools and preparation pipelines. Data cleaning, standardization, and transformation techniques are applied to ensure quality and consistency. Based on this, AI algorithms—especially Machine Learning—are trained to identify historical patterns, forecast demand for electronic meters, estimate manufacturing times, and detect potential operational bottlenecks. The third stage involves the development of an operational optimization module, responsible for applying mathematical optimization methods—such as Linear Programming, heuristics, and Genetic

Algorithms—to generate efficient production plans. The module considers multiple constraints, including line capacity, material availability, deadlines, scheduled maintenance, and specific requirements of Electronic Meters. The result is an optimized plan that sequences operations, balances workloads, and minimizes idle times. The fourth stage consists of implementing the Automated System, which integrates AI forecasts and optimized plans directly into the production line. This system sends automatic instructions to machines, monitors execution in real time, and continuously compares actual performance with planned targets. In the event of deviations—such as failures, delays, or material shortages—the system triggers a dynamic rescheduling mechanism [39-44]. Finally, the fifth stage involves the visualization and analysis of results, where intelligent dashboards present key performance indicators such as productivity, efficiency, lead time, and reliability. These insights enable continuous adjustments and foster a data-driven improvement culture. Thus, the computational methodology integrates AI, automation, and connectivity, enhancing the level of Operational Optimization in the production of Electronic Meters and aligning the process with Industry 4.0 principles.

III.1 INTENSITY OF THE TECHNOLOGICAL CHALLENGE

The research of this Development project involves uncertainties and challenges, which bring to light the original research premise with the objective of acquiring technical and scientific knowledge to implement tangible solutions. Due to these challenges, the project seeks to:

- Establish and manage the frequency and memory control of traceability devices and their processing capacity related to measurements, in order to maintain a stable state of data transmission and reception (Stages 1 and 2 of Ordinance 2,091);
- Identify patterns related to performance and quality indices to develop KPI control mechanisms and provide optimized decision-making recommendations (Stages 3 and 4 of Ordinance 2,091);
- Perform integrations with the hardware and software stack with minimal data loss and maximum quality in event synchronization to feed Business Intelligence systems (Stages 1 and 2 of Ordinance 2,091);
- Mitigate issues related to scrap, meter nonconformities, and dynamic changes in the production routing based on qualitative charts that generate insights for new solutions (Stages 3 and 4 of Ordinance 2,091).

The project proposes as a solution the development of an intelligent system capable of learning patterns based on events monitored and reported by testing devices, production routing and rescheduling mechanisms focused on electronic meters, and dashboards to support decision-making. The technical aspects of this project consolidate compliance with Ordinance No. 9,835/2022, Annex 1 (Intensity of the Technological Challenge):

I. **Level 4**: Conducting applied research — original research carried out with the objective of acquiring knowledge, primarily directed toward a specific practical objective or target;
II. **Level 3**: Conducting experimental development — systematic work based on pre-existing knowledge acquired through research or practical experience, aimed at producing new products and processes or improving existing ones.

III.2 SOLUTION STRUCTURING

The project includes technical constraints and limitations, making it essential to prioritize testing and validation in order to prove its effectiveness, in accordance with Ordinance No. 9,835/2022, Annex 1 (Solution Structuring).

I. **Solution structuring includes** identifying technical constraints or limitations for the solution and carrying out systematic development with validation tests. The limitations and restrictions are related to technical and security aspects that prevent certain processes, such as:

- **Monitoring area**: The project foresees the development of a traceability and data collection device with RS232 serial communication and via HTTP, TCP, and SSH. However, the monitoring area must be delimited according to the project scope;
- **Development sprints**: The project plans to adopt the technology in a localized space to control production-related events. However, due to time and cost constraints, development sprints were restricted to production control, traceability, quality, and monitoring;
- **Software and firmware upgrades**: The project includes materials and devices to meet its objectives; however, limitations regarding software and firmware upgrades were restricted to current data-processing technologies, which will be documented and delivered to the investing company.

To address these limitations, it is necessary to develop a system that monitors and generates indicators as a preventive premise against the project's technical limitations, highlighting innovation in the context of production rescheduling. For this purpose, several Industry 4.0 pillars will be implemented and incorporated, namely:

- **Data analytics**: To optimize decision-making processes and enable the system to inform and alert about possible harmful causes and effects of nonconforming process events, the project must implement data analysis measures using PyTorch and TensorFlow in Python, within a robust backend hosted on a server that provides adequate hardware for a well-defined Business Intelligence system. Simulation runs must be exhaustive in order to identify and define the best training architecture, thus delivering a viable computational model capable of supporting fluid and objective decision-making;
- **Systems integration**: Due to the business rules of each project module, it is necessary to separate responsibilities (Business Intelligence) through a microservices-based architectural model. For this purpose, stack definitions should focus on testing and simulating scenarios using Kubernetes, Docker, and/or other viable tools to minimize data-processing and systems-integration impacts, consolidating Stage 2 of Ordinance 2,091;
- **Artificial Intelligence**: AI will be used to perform production rescheduling with the premise of capturing causes and effects within the production system, enabling a proactive response to potential problems;
- **Internet of Things (IoT)**: IoT will be incorporated to collect real-time data from connected devices, providing a comprehensive view of the system and enabling dynamic adjustments as needed;
- **Digitalization**: Process digitalization will be a priority to ensure that all project stages are integrated and operate efficiently, enabling more effective management and a rapid response to any challenges that may arise.

III.3 SOLUTION STRUCTURING

The project includes technical constraints and limitations, and therefore testing and validation must be prioritized in order to demonstrate its effectiveness, in accordance with Ordinance No. 9,835/2022, Annex 1 (Solution Structuring).

I. **Solution structuring includes** identifying technical constraints or limitations for the solution and carrying out systematic development with validation testing. The limitations and restrictions are related to technical and security aspects that prevent certain processes, such as:

- **Monitoring area**: The project foresees the development of a traceability and data collection device with RS232 serial communication and via HTTP, TCP, and SSH. However, the monitoring area must be delimited according to the scope of the project;
- **Development sprints**: The project plans to adopt the technology in a localized space to control production-related events. However, due to time and cost limitations, development sprints were restricted to production control, traceability, quality, and monitoring;
- **Software and firmware upgrades**: The project includes materials and devices to meet its objectives; however, limitations regarding software and firmware upgrades were restricted to current data-processing technologies, which will be documented and delivered to the investing company.

To address these limitations, it is necessary to develop a system that monitors and generates indicators as a preventive premise against the technical limitations of the project, highlighting innovation in the context of production rescheduling. For this purpose, several Industry 4.0 pillars will be implemented and incorporated, namely:

- **Data analytics**: To optimize decision-making processes and enable the system to inform and alert about possible harmful causes and effects of nonconforming process events, the project must implement data analysis measures using PyTorch and TensorFlow in Python, within a robust backend hosted on a server that provides adequate hardware for a well-defined Business Intelligence system. Simulation runs must be exhaustive in order to identify and define the best training architecture, thus delivering a viable computational model capable of supporting fluid and objective decision-making;
- **Systems integration**: Due to the business rules of each project module, it is necessary to separate responsibilities (Business Intelligence) through a microservices-based architectural model. For this purpose, stack definitions should focus on testing and simulating scenarios using Kubernetes, Docker, and/or other viable tools to minimize data-processing and systems-integration impacts, consolidating Stage 2 of Ordinance 2,091;

- **Artificial Intelligence**: AI will be used to perform production rescheduling with the premise of capturing causes and effects within the production system, enabling a proactive response to potential problems;
- **Internet of Things (IoT)**: IoT will be incorporated to collect real-time data from connected devices, providing a comprehensive view of the system and enabling dynamic adjustments as needed;
- **Digitalization**: Process digitalization will be a priority to ensure that all project stages are integrated and operate efficiently, enabling more effective management and a rapid response to any challenges that may arise.

III.4 CURRENT STATUS OF THE SYSTEM

Considering the company's current scenario, there is a Lean Intelligent Manufacturing System (SIME) in place, in which a system foundation has already been developed through a data persistence layer (DBMS) using a MySQL database hosted on a central server. There is also a business layer that uses a set of tools to optimize request processing, such as FastAPI, Celery, WebSockets, among others, all of which use Python as the main programming language, creating a DevOps environment and architecture for communication with the system. Finally, there is a presentation layer that visually displays some production-related raw data, such as the number of meters produced, production targets, and overall line production rates. However, this information is not individualized by each piece of equipment, as it is accounted for through the meter sealing control module.

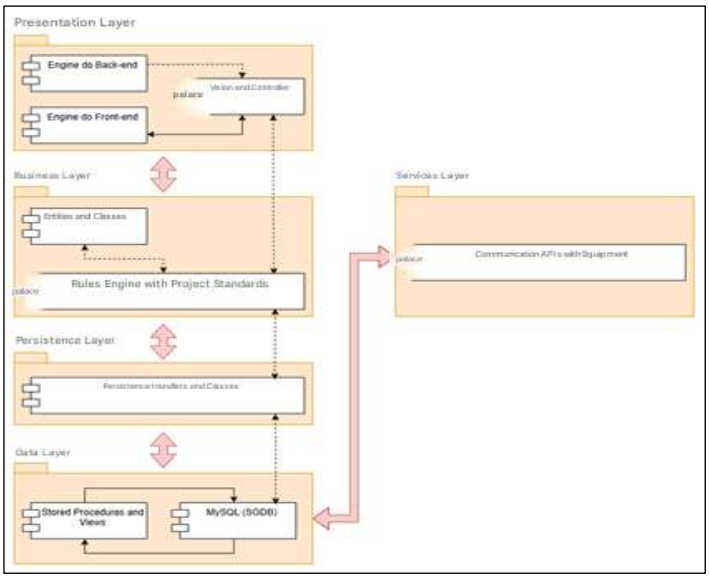


Figure 1: Current Status of the System.
Source: Authors, (2026).

Figure 1 presents a representation of these layers of the existing system, providing a visualization of what they represent and how they are interconnected. This is known as the Model-View-Controller (MVC) model, which revolves around the definition of design patterns that manage stored data, either for display in the presentation layer or for direct data editing through communication with the persistence layer. In summary, this structure was developed within the SIME software and encompasses user and permission management, as well as Business Intelligence through the HORUS API, which organizes data from two main modules: Meter Sealing Control and Inventory Control. These two modules allow monitoring of resource relationships and data on the number of meters produced, with a dedicated dashboard for this purpose. The current project aims to improve the SIME system to encompass the development of an intelligent system for production planning. This system is based on monitoring production events, which indicate actions to be taken to follow the planned schedule, integrating resource management and the Bill of Materials (B.O.M.), currently managed by the JUNDSOFT software, while optimizing business rules with new design patterns.

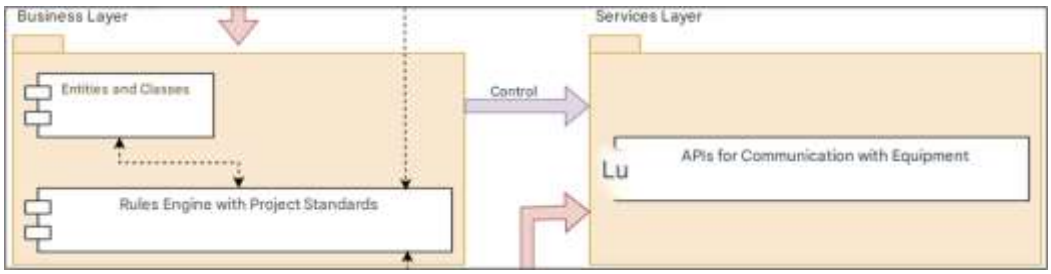


Figure 2: Business and Services Layers.
Source: Authors, (2026).

The current research aims to improve the SIME system to encompass the development of an intelligent system for production planning. This system is based on monitoring production events, which indicate actions to be taken to follow the planned schedule, integrating resource management and the Bill of Materials (B.O.M.), currently handled by the JUNDSOFT software, while optimizing

business rules with new design patterns. As can be seen in Figure 2, the services layer contains the APIs developed for communication with the equipment and, therefore, continuously collects data from these devices directly into the database.

III.5 TRACEABILITY DEVICE

In this project, the objective is to integrate the software stacks (services layer API) with the hardware stacks (Traceability Device), which in turn are connected to equipment that generally use RS232 serial communication. The device will serve as the interface to communicate with the equipment and connect them to the internal network, advancing the equipment from a Stage 1 maturity level to Stage 2, where connectivity exists for data collection from each piece of equipment in a pilot line, which will be used to validate the data collection. Some devices in this pilot production line do not have connectivity and will receive a traceability device for data collection, as outlined below:

- **Parameterization Device**: This prototype device will be responsible for configuring the product's memory, recording the serial number, and allowing verification of other quality parameters to ensure that the product meets the customer's purchase order requirements (Verification according to the production order). This system must operate automatically, without operator intervention. Based on the production order number, the system will query via API which configurations and checks need to be performed on the product. Since these configurations are based on customer parameters, the device is called Parameterization.
- **PIMA Asynchronous Serial Output Device**: This device will be a functional test prototype for the product's communication protocol. It will ensure that the product functions and communicates correctly and that the serial code recorded in memory (done by the Parameterization Device) is correct. This stage occurs after the product is permanently closed, ensuring that there is no error in lid exchange or that the mechanical closing process did not damage the product. To determine what should be tested, the system will query via API, based on the production order number, the correct serial number in the production sequence, the communication protocol, and other optional items to be verified.

III.6 SYSTEM MODULES

With the collection of these data, which will be presented in detail in the project scope, it is possible to create new business rules in a new module called the **Production Control Module**, which is responsible for receiving data from each piece of equipment in real time and associating it with production events. These events represent higher-level information that allows for more assertive decision-making. Through this module, the project aims to develop intelligent system software for production planning. The **Traceability Module** is designed to organize data regarding product batches and their journey through the production line. This makes it possible to identify potential batch failures because, in the case of a defect in a meter during the final testing stage, it is possible to track the exact moment the product passed through each piece of equipment. This increases reliability in quality control processes, among other benefits of product traceability. These data allow the creation of business rules for updating the production planning based on traceability information. Meanwhile, the **Quality Control Module** collects, organizes, and makes visible the data related to product quality, allowing for a more comprehensive analysis of detected defects to define standards that will be used to adjust production planning and optimize the company's quality processes. Finally, the **Equipment Module** is not related to the products themselves but to the machines that produce them. Data from this module will be collected to observe patterns in equipment operation, enabling the scheduling of preventive maintenance and providing a basis for predictive maintenance. Additionally, with business rules associated with machine productivity, it is possible to reorganize production planning concerning machine health, operating hours, and reduce failure cycles. Each of these modules is a fundamental component for improving visibility and gathering information to determine production events, which contributed to the creation of an intelligent production planning system. For this, the data to be digitized from each module were defined, and from this data, relevant indicators will be constructed. These are high-level information derived from real-time data and are used to determine production events and shape production planning. The modules and their data and indicators are organized as follows:

- **Production Control:**

  - *Digitized Data:* Equipment utilization rate; production cycle time; number of products manufactured; machine downtime and corrective/preventive maintenance data; raw material consumption.
  - *Relevant Indicators:* Productivity rate; unplanned downtime index; raw material optimization index; reduction rate of losses and rework.

- **Traceability:**

  - *Digitized Data:* Complete batch history (dates, locations, and responsible personnel); detailed production step records; event logs and movements of each product/batch; return and defective product data.
  - *Relevant Indicators:* Real-time visibility index; tracking error reduction rate; response time for fault and return identification; efficiency in locating problematic or defective batches.

- **Quality Control:**

  - *Digitized Data:* Rejection rate per batch; nonconformities identified in inspections; test and quality assessment results; history of rework and repairs performed.
  - *Relevant Indicators:* Defect rate per batch; product compliance with quality standards; number of customer complaints; inspection and quality control time.

- **Equipment:**

  - *Digitized Data:* Equipment operating and downtime; frequency and type of maintenance performed (preventive/corrective); component and part lifespan; energy efficiency and resource consumption data.
  - *Relevant Indicators:* Mean time between failures (MTBF); mean time between corrective maintenance; operational efficiency of equipment; maintenance costs.

IV. RESULTADOS

Equipment, Quality Control, and Production

Initially, a review of the PUR was conducted, along with a study on non-connected devices to define the traceability system. The possibilities for integration with the SIME and JUNDSOFT software were evaluated to obtain metadata from the B.O.M. During this process, the operation of the **Parameterization Device** and the **Asynchronous Serial Output Device (PIMA)** was analyzed. The use of a **Raspberry Pi 5** for data collection and processing was considered, given its memory capacity and compatibility with multiple programming languages, which facilitate real-time solution implementation.

IV.1 PARAMETERIZATION DEVICE

The Parameterization Device was identified as a critical system for the automatic configuration of product memory. Its function is to ensure that each manufactured unit complies with the company's customer specifications, recording the serial number and verifying essential quality parameters. The operation of this device will be fully automated, eliminating the need for manual intervention. Integration via API will allow querying configurations and checks for each produced item, providing more precise and efficient quality control. This approach aims to eliminate human errors and ensure standardization in production, improving product safety and reliability.

IV.2 ASYNCHRONOUS SERIAL OUTPUT DEVICE (PIMA)

The PIMA device is responsible for performing functional tests on the product's communication protocol. This equipment ensures that the serial number recorded in memory is correct and that the product operates as expected. Verification occurs after the product is permanently sealed, preventing issues such as incorrect lid exchanges or mechanical damage during the closing process. The PIMA operation will also be automated, allowing it to automatically query production order information via API. This process ensures the accuracy of the production sequence, validates the communication protocol, and performs additional tests according to customer specifications. Implementing this device aims to increase production reliability and reduce operational failures.

IV.3 INTEGRATION AND DEVELOPMENT OF THE PRODUCTION CONTROL MODULE

From the data collected by these devices, the **Production Control Module** will be developed to consolidate real-time information and transform it into production events. These events enable a deeper analysis of production line performance, facilitating strategic decision-making and optimizing manufacturing management.The integration of these devices with the Production Control Module will enable:

- Real-time monitoring of parameterization and verification processes;
- Automatic alerts for deviations from quality standards;
- Equipment performance analysis and identification of potential production bottlenecks;
- Creation of a traceability history for each manufactured unit.

Integrating the Parameterization and PIMA devices into the new Production Control Module will provide significant benefits for production management, increasing efficiency and reducing operational failures. The use of automated technologies and API integration allows for a more robust approach to quality control and process standardization. During the analyses, it was identified that if mechanical adjustments are needed for device integration into the production environment, **CAD (Computer-Aided Design)** modeling will be used. CAD allows the creation, modification, analysis, and optimization of 3D digital projects, providing precision and efficiency in developing mechanical components.

For this purpose, **SolidWorks** software will be used, one of the most advanced tools for parametric modeling and mechanical simulation. SolidWorks enables detailed part and assembly design, structural analysis, and fit testing before manufacturing. Using this technology will facilitate the creation of supports, enclosures, and other adaptations required to install the devices in the industrial environment, ensuring better performance and integration with the systems.

IV.4 CONNECTIVITY AND NETWORK

A microcomputer kit was considered to process data directly on the devices and communicate with the systems being developed. The **Raspberry Pi 5** is a compact and versatile microcomputer with 8GB of RAM. As shown in the following figure, it is suitable for projects requiring real-time data processing and storage. This model is particularly appropriate to support the development of the Parameterization and PIMA prototypes, as it allows local data processing and the execution of parameterization and production analysis software.



Figure 3: Raspberry Pi 5 8GB.
Source: Authors, (2026).

It has high memory capacity, and its compatibility with multiple programming languages will allow the creation of a robust and agile testing environment, in addition to facilitating integration with sensors and devices. In this context, the **Raspberry Pi 5** will play a highly relevant role in the communication between devices and the software developed for the project. The research conducted during this period highlighted the main contributions, including:

- **Communication Interface:** Can act as a communication gateway, connecting the Parameterization and PIMA devices to the Production Control Module;
- **Data Collection and Processing:** Allows real-time acquisition of data generated by the devices, ensuring integrity and efficient transmission to the analysis systems;
- **Integration via Industrial Protocols:** Supports protocols such as MQTT, Modbus, and OPC UA, enabling interoperability between different equipment and the system;
- **Automation and Control:** Can be used to automate commands, ensuring devices operate according to the guidelines established by the control software;
- **API Connectivity:** Facilitates integration of devices with system APIs, allowing dynamic queries and real-time adjustments to production settings.

The implementation of the Raspberry Pi 5 will enhance system efficiency and reliability, reducing response times and optimizing device management on the production line. This period of activities also included a review of the **Connectivity and Network** point, as described in the PUR assumptions, due to the following condition:

- **Connectivity and Network:** Assumes that the SIME system execution environment, as well as the monitored machines, have the capacity to be connected to SIME through a local network (LAN), either directly or via an intermediate IoT prototype (microcomputers, e.g., Raspberry Pi) and sensors for data collection.

Analyzing this assumption was important to assess the feasibility of communication between the SIPE system and the monitored machines. This initial study helped identify potential technical limitations, such as network infrastructure constraints, and define strategies to ensure effective device connection. Additionally, it enabled the selection of the best approach for data transmission through intermediate devices like the Raspberry Pi, which assist in collecting and sending real-time information. With efficient connectivity, **SIPE** will be able to consolidate operational data reliably, enabling precise analyses to optimize decision-making. Implementing sensors and intermediate devices like the Raspberry Pi will allow the integration of devices connected to the system, ensuring that all equipment can communicate with the production management platform. Furthermore, compatibility with industrial protocols such as MQTT, Modbus, and OPC UA will allow standardized and secure communication between the system's different components. During this project period, research also focused on analyzing frameworks, libraries, and technological protocols essential for system development. Solutions for Front-End and UI/UX development, inter-system communication, database management, and real-time data processing were evaluated. This study identified the most efficient tools to meet project requirements, ensuring technical and operational feasibility, as summarized below.

a) **Front End and UI/UX:**

- **Vue.js:** JavaScript framework for creating dynamic and modular user interfaces;
- **Vue Router:** Library for managing routes in Vue.js applications;
- **Vue I18n:** Plugin for internationalization and localization of applications;
- **VueUse:** Collection of utilities for interacting with APIs and sensors;
- **Quasar/Naive UI:** Component frameworks for developing web, mobile, and desktop applications.

b) **Data Visualization:**

- **ECharts:** Library for creating interactive and dynamic charts;
- **Moment.js:** Library for date manipulation and formatting.

c) **Front End and Back End Communication:**

- **Axios:** HTTP client for asynchronous communication with RESTful APIs.

d) **Back End and Database:**

- **FastAPI:** Framework for building fast and secure APIs, with support for static typing;
- **SQLAlchemy:** ORM library for efficient interaction with SQL databases;
- **Redis:** In-memory database for caching and processing large volumes of data;
- **Celery:** Library for executing asynchronous tasks and distributed processing.

e) **Communication Protocols:**

- **WebSockets:** Protocol for real-time bidirectional communication.

f) **Validation and Testing:**

- **Pydantic:** Library for data validation based on Python typing;
- **Pytest:** Framework for unit, integration, and end-to-end testing.

g) **Internet of Things (IoT):**

- **IoT:** Technology for interconnecting physical devices, enabling real-time data collection and analysis, applied in parameterization and PIMA.

A careful analysis of technologies, tools, and frameworks enabled the selection of solutions suitable for an efficient and structured implementation. The combined use of these technologies will ensure a modular, secure, and high-performance system aligned with the project's objectives. Subsequently, the integrated systems development team focused on researching traceability devices—particularly the Parameterization and Asynchronous Serial Output (PIMA) devices—and structuring the embedded architecture with the Raspberry Pi 5 as the central processing and communication hub. The IoT Developer, Cássia Gabrielly de Souza Lopes, actively participated in the technical review of the PUR, contributing to the validation of hardware components, with emphasis on the Raspberry Pi 5, considered a key element in the collection, processing, and transmission of data from the devices. The researcher conducted detailed analyses of the embedded board's technical specifications, such as its ARM Cortex-A76 processor, 8GB RAM, GPIO interfaces, and support for industrial protocols (MQTT, OPC UA, Modbus), highlighting its potential for integration with the management system and connected devices. She also assessed the feasibility of using serial and USB interfaces for direct communication with production and testing devices. The Raspberry Pi 5 was identified as an essential component for the system architecture, playing a central role in communication between devices and the developed software. The research highlighted its main contributions, including its ability to act as a communication gateway connecting the devices to the Production Control Module, enable real-time data acquisition and processing, integrate via industrial protocols such as MQTT and HTTP, automate operational commands, and provide connectivity with system APIs for dynamic queries and real-time adjustments. Implementing this device will significantly enhance system efficiency and reliability, optimizing device management on the production line. Research was also conducted on the embedded board's electrical diagram, as shown in Figure 4.

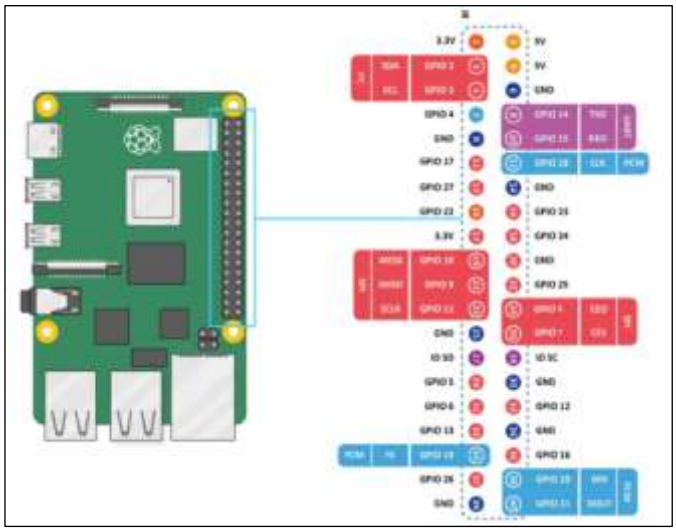


Figure 4: Raspberry Pi5.
Source: Authors, (2026).

This embedded model is the most advanced version of the popular single-board computer line developed by the Raspberry Pi Foundation. With significant improvements in performance and connectivity, the device provides a robust solution for industrial applications, automation, and IoT. Its main technical specifications are as follows:

- **Processor:** Quad-core ARM Cortex-A76, with a frequency of up to 2.4 GHz, delivering performance approximately three times higher than its predecessor, the Raspberry Pi 4.
- **RAM:** Available in 4 GB and 8 GB LPDDR4X-4267 versions, allowing greater energy efficiency and improved performance for multitasking applications.
- **Storage:** MicroSD UHS-I slot, with additional support for storage via PCIe 2.0 (through a dedicated connector), enabling the use of NVMe SSDs for higher read and write speeds.
- **Connectivity:** Includes two USB 3.0 ports and two USB 2.0 ports, plus a Gigabit Ethernet port that supports Power over Ethernet (PoE) when used with an additional module.
- **Graphic Interfaces:** Equipped with two micro-HDMI 2.0 outputs, supporting 4K resolution at 60 Hz and HDR compatibility, ideal for applications requiring high-quality graphics.
- **Graphics Processing:** VideoCore VII GPU with support for OpenGL ES 3.1, Vulkan 1.2, and hardware-accelerated video decoding for H.265 (4Kp60) and H.264 (1080p60).
- **Wireless Connectivity:** Supports Wi-Fi 5 (802.11ac) and Bluetooth 5.0, ensuring stable connections for wireless networks and Bluetooth devices.
- **PCIe Bus:** Integrates a PCIe 2.0 connector, allowing the addition of high-performance peripherals, such as storage controllers or expansion modules.
- **GPIO Pins:** Features 40 GPIO pins compatible with previous Raspberry Pi versions, enabling connections to sensors, actuators, and other devices.
- **Power Supply:** Uses a USB-C connector for power input, supporting 5V/5A, providing sufficient power to operate additional peripherals.

To ensure efficient communication of the traceability and functional testing devices, the Raspberry Pi 5 will be used as an intermediary between the devices and the management system. The Parametrization Device will be connected to the Raspberry Pi via USB or GPIO, allowing automated configuration of the product memory and recording of the serial number. Through API queries, the Raspberry Pi will receive the appropriate configurations and transmit them to the Parametrization Device. The PIMA Asynchronous Serial Output Device will also be integrated with the Raspberry Pi 5, which will serve as a validator of the product's functionality. The Raspberry Pi will communicate with the PIMA via serial protocol, verifying the correct recording of the serial number and checking the device's functional communication. In case of inconsistencies, it will be able to log faults and alert the production system in real time. The study identified technical limitations related to network infrastructure and defined strategies to ensure efficient device connectivity.

Compatibility with industrial protocols such as MQTT, Modbus, and OPC UA will ensure standardized and secure communication between system components. During the analyses conducted, no mechanical adjustments were identified as necessary for integrating the devices into the production environment. Therefore, CAD modeling—which allows for the creation and optimization of three-dimensional projects—was not required at this stage. To Marcelo Maia do Nascimento, a senior embedded hardware and software developer, collaborated with the IoT Developer in developing the functional prototype for reading data from electronic meters via the national PIMA protocol. He conducted a comprehensive technical study of PIMA, covering its packet structure (preamble, identifier, size, scope + index, data, and CRC) as well as ANEEL regulatory requirements for interoperability across PLC, RF, and RS-485 networks. He developed the communication logic between the meter and the microcontroller through serial terminals (MTx and MC), with subsequent data transmission to a WebServer embedded in the Raspberry Pi 5, accessible remotely via web browsers.

IV.5 STUDY OF THE PIMA PROTOCOL AND APPLICATION

The Developer conducted an in-depth study of the PIMA protocol, created in Brazil to ensure interoperability among meters from different manufacturers and management systems. The protocol was analyzed in its entirety, from operation over media such as PLC, RF, and RS-485 to its data packet structure, including preamble, identifier, size, scope + index, data, and CRC fields. This deep understanding allowed technical implementations to comply with the 2400 bps speed parameter and the unidirectional asynchronous transmission structure of electronic meters, according to ANEEL specifications.

IV.5.1 Development of the Serial Communication Prototype with WebServer

It will be necessary to adapt the PIMA device, based on the Raspberry Pi 5 microcontroller, to read data sent via the unidirectional serial output of the meters (MTx and MC terminals). Communication between the meter and the microcontroller will be established using physical jumpers, enabling the capture of transmitted packets. The Raspberry Pi 5, with native Wi-Fi connectivity, can be programmed to communicate with a local WebServer, which will display the measurement information in an interface accessible remotely via a browser. The workflow is illustrated in Figure 5.



Figure 5: Serial Communication Prototype with WebServer.
Source: Authors, (2026).

IV.5.2 Data Conversion and Visualization

During the research, it was observed that the structure of the PIMA packet is decoded to extract the data of interest, such as the meter serial number, voltage, current, and power parameters, among others. After reading and decoding, the data will be organized and sent to the WebServer, which will display the values in real time for the operator. This functionality will enable continuous monitoring of energy consumption, promoting greater awareness and control by the consumer. Figure 6 shows the simplified electrical diagram of unidirectional asynchronous serial communication between the meter and the remote device (PIMA Device). On the left side, the serial outputs of the meter are represented, while on the right side, the serial input of the receiving device can be observed. Although the components may vary according to the manufacturer, the basic connection configuration was maintained in this project: the MRx pin of the PIMA device is connected to the MTx pin of the meter, while the MC is grounded. When the meter transmits data through the MTx pin, the device receives the information packets, processes the data, and makes them available on a WebServer developed for visualization. This implementation ensures compatibility with the PIMA standard, allowing proper reading and display of the measurement information.

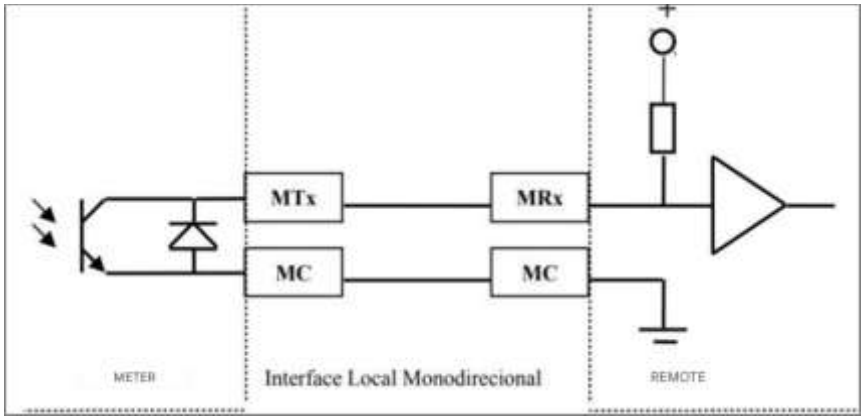


Figure 6: PIMA Electrical Diagram.
Source: Authors, (2026).

The formation of the data packets sent through the asynchronous serial output must follow the specific formatting described below:

- **Preamble:** Initial signaling of a packet. It consists of 2 bytes with the hexadecimal characters AA and 55;
- **Identifier – Meter serial number:** Composed of 5 bytes in BCD format, allowing a 10-digit numbering;
- **Length:** Count of the number of bytes related to the SCOPE + INDEX characters and DATA. Composed of 1 byte;
- **Scope + Index:** Identifier of the type of information to be sent, according to the standard defined in Table 1;
- **Data:** Corresponds to the actual measurement values, presented in BCD format. These data must comply with the definitions of the Application Protocol established by the Brazilian Association of Technical Standards (ABNT);
- **CRC:** Refers to the cyclic redundancy check character of the CRC16 message ($X^{16} + X^{15} + X^2 + 1$), applied over all bytes of the packet, except for the PREAMBLE and the CRC itself.

Automated communication with the meters will enable detailed traceability of the produced units and real-time validation of readings, ensuring greater safety and efficiency on the production line. The development enabled the creation of a continuous flow for reading, decoding, and visualizing meter data, validating the serial number, electrical parameters, and the overall operation of the product. The embedded architecture ensures real-time communication, with the possibility of integration with the SIPE system through APIs, promoting full traceability of the produced units. During the period under analysis, significant hours were dedicated to conducting in-depth technical studies aimed at directly assisting the technical coordinator and the hardware and IoT developers in refining the functional and operational requirements of the system under development.

The author's contribution was essential in consolidating the understanding of demands related to parameterization and communication with electronic energy meters, ensuring that the proposed solutions were aligned with the technical standards of the Brazilian electrical sector, especially with regard to the implementation of connectivity in the PARAMETERIZATION device. In addition, the researcher contributed to defining strategies for integration between physical devices and the embedded and web software architecture, providing a solid foundation for the development and validation of the prototype. The Parameterization Device is an industrial piece of equipment used in the process of configuring and writing parameters to electronic energy meters during the manufacturing phase or final calibration. It plays a fundamental role in customizing meters according to customer technical specifications or current regulatory requirements (e.g., ANEEL, INMETRO, ABNT).

IV.5.3 Objectives and Main Functions

The main objective of the device is to automate the writing and verification of the meter's internal parameters, ensuring its identity, functionality, and compliance with technical standards. Among the functions adopted at Wasion are:

- Writing the serial number to the meter (unique identifier).
- Configuration of tariff tables and calendars.
- Parameterization of demand limits, load profiles, and events.
- Enabling communication modes (e.g., PIMA, DLMS/COSEM).
- Writing technical data such as nominal voltages, base currents, and frequencies.
- Validation and verification of embedded firmware.
- Updating cryptographic keys in meters with advanced security functions.
- Integration with traceability systems and production order databases.

a) Architecture of the PARAMETERIZATION Device

The architecture of the parameterization device includes:

- Central Controller (Arduino Mega microcontroller) with dedicated software.
- Programmable power supply to energize the meter according to its nominal parameters.
- Read and Write Module (via serial port).
- Cryptographic Security Module (in meters with authentication).
- Computer application for configuration and visualization.
- Local database to store parameters and operation logs.

b) Communication Technologies

In discussions with the Technical Team, a study was conducted to identify potential connectivity applications for the device:

- IEC 62056-21 standard optical port: very common in meter parameterization.
- RS-232 or RS-485 serial interface: mainly used for bench parameterization.
- USB port with optical converters: allows connection with parameterization software.
- Wi-Fi / Ethernet interface: in more modern devices with remote control.
- Protocol standards: the most common are PIMA (in Brazil) and DLMS/COSEM (international).

c) Application of the PARAMETERIZATION Device at Wasion

It was identified that Wasion has flexibility in its production lines, which are fully customizable to adapt manufacturing processes as needed. Regardless of the production line, the parameterization device is used at specific stages:

Post-assembly / Pre-closing

- Writing production and identification data.
- Verification of basic operation.

Post-calibration

- Parameterization based on the results of metrological adjustment.

d) Integration with Production Systems

e) Currently, the parameterization device does not have integration with:

- MES (Manufacturing Execution System): enables traceability of the produced unit.
- ERP: for consultation of production orders and customer requirements.
- SQL / NoSQL databases: where parameters are stored and versioned.

Automation/connectivity of the device will minimize human errors, increase productivity, and ensure technical compliance, being a mandatory step in the metrological certification of meters by INMETRO.

a) Standards and Regulations

b) The operation and parameterization of meters follow standards and regulations, among which the following stand out:

- INMETRO Ordinance No. 457/2021: regarding legal metrology of meters.
- ANEEL Resolution No. 844/2019: establishes guidelines for metering and quality.
- ABNT NBR 14522 Standard: specific characteristics of electronic meters.
- IEC 62052 and IEC 62056 Standards: address communication protocols and general requirements.

In the context of the research, the Parameterization Device will be used to:

- Automate the writing of serial numbers and technical parameters of meters.
- Integrate with the traceability module, sending data to an internal WebServer.
- Automatically verify whether the meter complies with the production order specification.
- Record data in an internal database for quality control and statistical analysis.

During this period, the team also assessed the need for mechanical adjustments to the devices, with planned use of SolidWorks software for modeling supports and enclosures. Integration with the SIME and JUNDSOFT systems was considered in defining database structures and communication protocols, establishing a foundation for the development of traceability, production control, and quality control modules. The result of this activity was the consolidation of a robust technical infrastructure for the development of embedded devices, establishing the foundations for functional testing, real-time data collection, and integration with the production management systems of the investing company. During the current month, activities related to the Integrated Systems Development stage advanced significantly, with the direct involvement of the IoT Developer, the Technical Technology Specialist, and the Senior Embedded Hardware and Software Developer. The main focus was on the technical enhancement of the Parameterization Device for electronic meters, a strategic element of the SIPE system to ensure traceability, quality control, and customization of devices according to Wasion's requirements.

The IoT Developer focused efforts on creating two fundamental diagrams for updating the parameterization device, based on the technical and structural studies carried out in the previous month. From a detailed analysis of the functions, communication ports, and capabilities of the Raspberry Pi 5 microcontroller, the electronic components required to ensure proper communication between physical devices and the production management system were defined. The diagrams were developed prioritizing functional logic compatible with the existing structure of the equipment, respecting component datasheets and carefully distinguishing electrical circuits from electronic circuits. This distinction was essential to ensure that interconnections between signal and power elements were designed safely and accurately. The first diagram, an electrical circuit modeled using Proteus software, focused on integrating the Raspberry Pi 5 into the parameterization device circuit. This simulation was essential to validate interactions among components before physical construction.

In the designed circuit, the Raspberry acts as the main interface, receiving configuration parameters via API and transmitting this information to the meter. Communication between the embedded system and the meter was enabled through a USB-to-TTL adapter, which connects the Raspberry's serial ports to the pins responsible for reading and writing data on the meter. The device interface was also designed to display the parameterization status to the operator, providing real-time visibility and greater control over the process. Before practical assembly, all connections were meticulously simulated in Proteus, allowing the anticipation of failures such as short circuits, voltage incompatibilities, and risks of damage to sensitive components. This virtual validation ensured the integrity and safety of the circuit. After this phase, physical assembly began in a segmented manner, enabling modular testing at each stage, which increased efficiency in fault identification and contributed to a more stable and secure final integration.

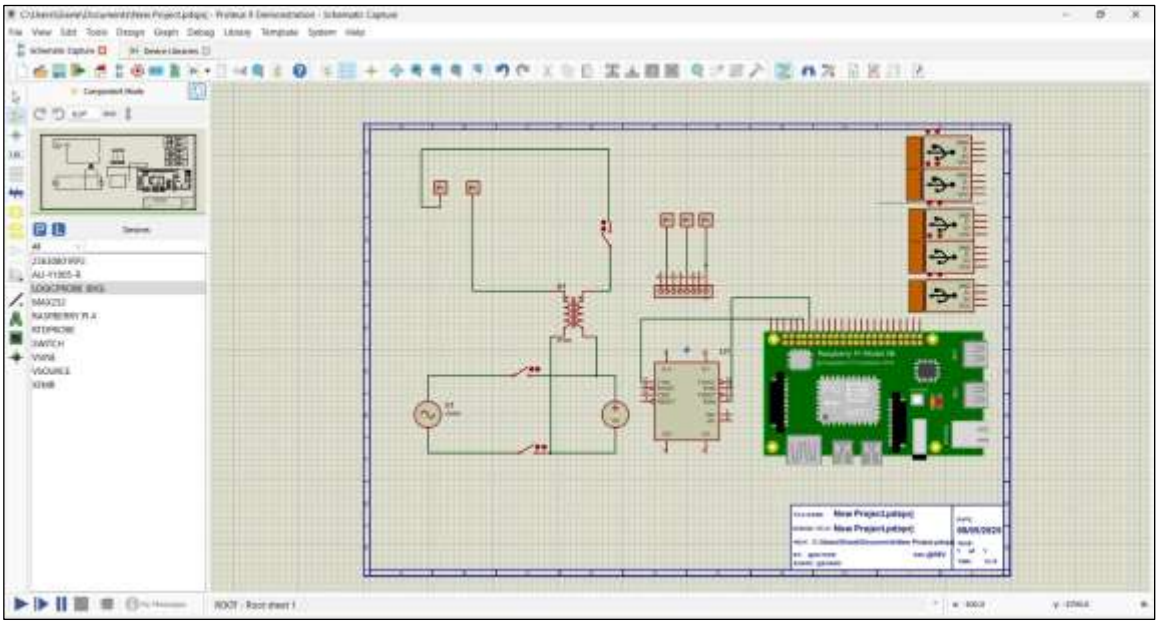


Figure 7: Electrical diagram of the parameterization device.
Source: Authors, (2026).

In parallel, the research deepened its investigation into the company's production structure, mapping with greater precision the stages of the assembly line in which the parameterization device is inserted. This analysis was formalized through technical flowcharts, as presented in the figure below, and enabled a more detailed understanding of the critical integration points between the equipment and production flows, especially at the pre-closing stage of the meters, known as the Ultrasonic Station. As a result of this mapping, meetings were held with the company's team to discuss the compatibility of communication protocols with the manufactured meters. In these meetings, emphasis was placed on the technical and regulatory feasibility of implementing the PIMA protocol, considering the requirements of ANEEL, INMETRO, ABNT, and IEC.

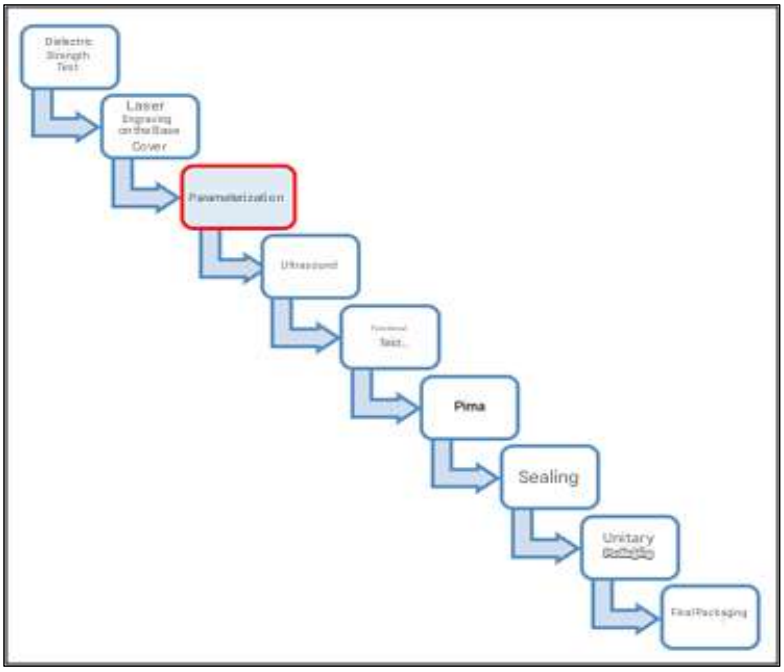


Figure 1 Process Flowchart for Single-Phase Meters.
Source: Authors, (2026).

Simultaneously, the Senior Embedded Hardware and Software Developer directed efforts toward advancing the technical solution of the parameterization device. Initially, the technical and functional requirements of the prototype were defined, with a focus on compatibility with the communication standards adopted by Wasion. Acting as an operational interface, the device was designed to allow operators to configure meters in an automated, secure, and traceable manner. Parameters such as serial number, calibration, firmware version, and operational data were defined as essential to the process. The Raspberry Pi 5 was retained as the core of the system, and logical level converters, USB-to-Serial adapters, and EEPROM modules were incorporated into the project, which are necessary for simulations of writing, reading, and integrity verification. The test circuit was initially modeled using Draw.io. The device firmware was structured using Python, with libraries such as *pyserial* for serial communication and *smbus2* for I²C integration. The initial routines focused on reading unique identifiers, secure write commands, and subsequent validation through reading, ensuring robustness even under conditions of noise or protocol variations, as shown in Figure 9.

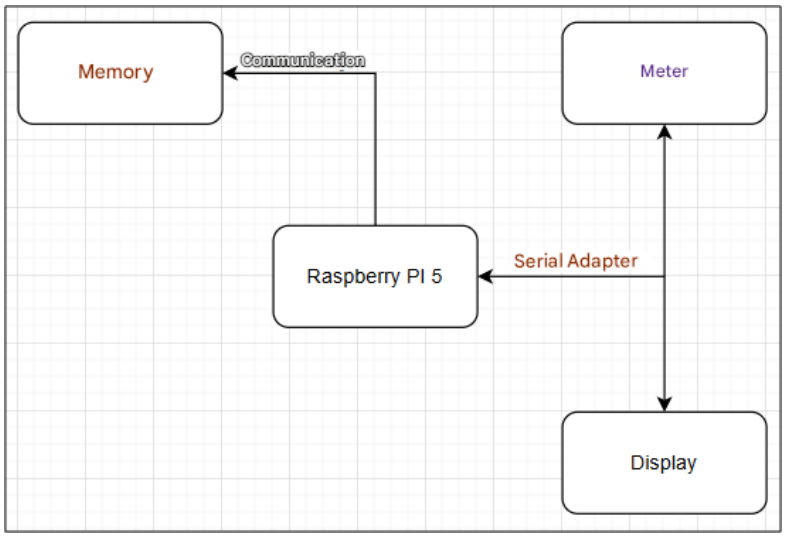


Figure 9: Diagrama para o circuito de teste.
Source: Authors, (2026).

Additionally, an in-depth analysis of the electrical and logical interconnection points was carried out, considering critical aspects such as:

- Power supply compatible with the factory infrastructure (12 V, 24 V, or 220 V AC/DC).
- Shielding against electromagnetic interference, which is common in industrial machine environments.
- Robust and reliable communication interfaces, using standards such as RS-485, Ethernet, MQTT, Modbus, and communication via REST APIs.

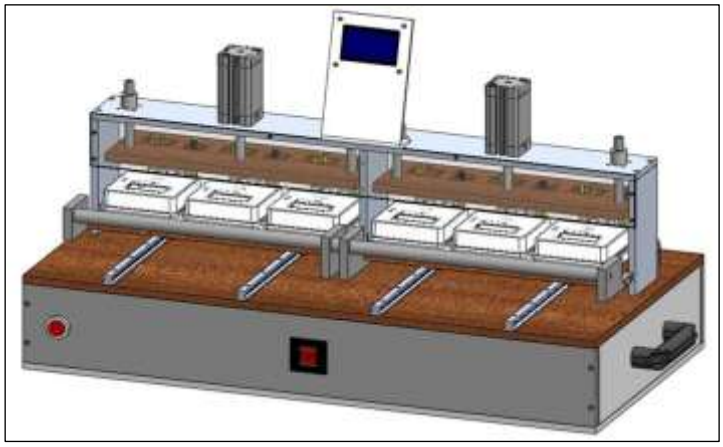


Figure 10: 3D Model of the PIMA Verification Device.
Source: Authors, (2026).

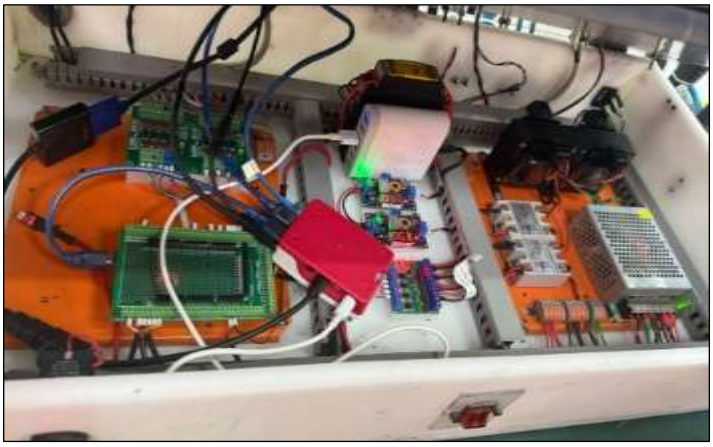


Figure 11: Hardware of the PIMA Verification Device.
Source: Authors, (2026).

Complementing this architecture, the use of Apache Kafka was defined as a distributed messaging platform, serving as the core of data communication between the physical devices (Parameterization and PIMA Device) and the backend systems. Kafka enables real-time event processing, offering (Figure 12):

- High fault tolerance.
- Unlimited horizontal scalability, with the possibility of expanding brokers as data volume grows.
- Long-term data persistence, ensuring reliability in the transport of critical production messages and events.

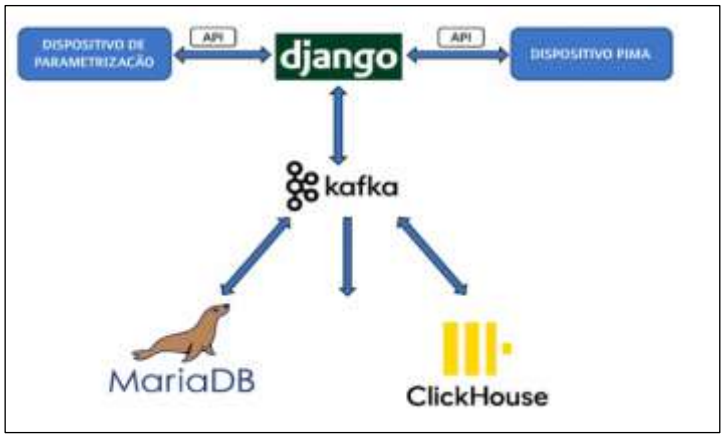


Figure 12: Backend and Communication Architecture Diagram.
Source: Authors, (2026).

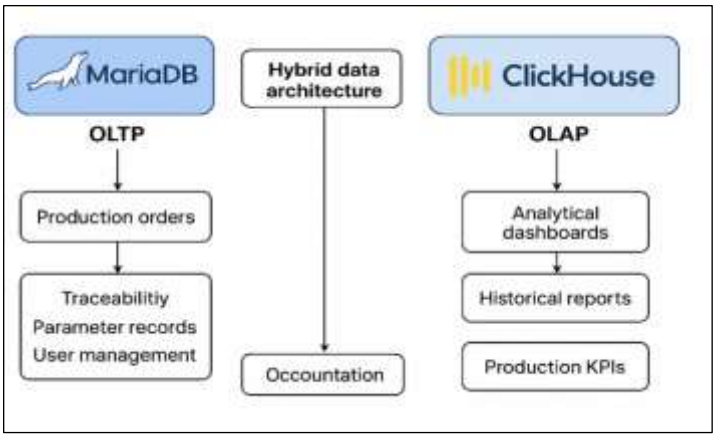


Figure 13: Hybrid architecture for data storage.
Source: Authors, (2026).

Regarding data storage, a hybrid architecture was consolidated, combining two databases with distinct purposes:

- **MariaDB**, adopted for OLTP (Online Transaction Processing), is responsible for managing operational data such as production orders, traceability, parameterization records, and user control. Its selection was based on robustness, relational integrity, scalability, and wide adoption in industrial environments.
- **ClickHouse**, adopted for OLAP (Online Analytical Processing), is intended for managing analytical data, allowing the processing of billions of rows with extremely high-speed queries. Its columnar architecture is ideal for generating analytical dashboards, historical reports, production performance monitoring, and KPIs such as OEE (Overall Equipment Effectiveness) and FPY (First Pass Yield).

MongoDB was also considered during the evaluation process, mainly for storing semi-structured data. However, given the specific project requirements—which demand high transactional consistency and large-scale analytical performance—the technical decision prioritized the combination of MariaDB and ClickHouse, providing robustness and superior performance. Figure 13 clearly and objectively illustrates the hybrid data storage architecture adopted in the project, combining MariaDB (OLTP) and ClickHouse (OLAP), each performing specific and complementary roles within the ecosystem of the developed solution.

IV.5.4 Frontend Development and Operational Interface

The frontend team, composed of Leonardo and Matheus, made significant progress in defining the visual architecture and prototyping system screens, highlighting:

- **Adopted Framework:**

  - Vue.js, combined with Tailwind CSS for styling and Vite for build optimization and agile development.

During subsequent technical meetings (Figures 14, 15, and 16), the use of the Vue.js framework was defined as the basis for system interface development. The choice considered its lightweight nature, ease of integration with WebSocket, modularity, and accessible learning curve, facilitating collaboration among team members. Vue.js will enable the creation of reactive components for real-time data visualization, with reduced complexity in maintenance and scalability of the application.



Figure 14: Vue.js.



Figure 15: Tailwindcss.
Source: Authors, (2026).



Figure 16: Vite.
Source: Authors, (2026).

A Simplified Architecture Diagram was created to clearly, objectively, and visually represent the structure of the system for parameterization and segregation of electronic energy meters. The main goal of this representation was to provide a high-level view of the system's key components, their functional responsibilities, and the communication flows between software modules, embedded hardware, and external backend systems, as shown in Figure 17.

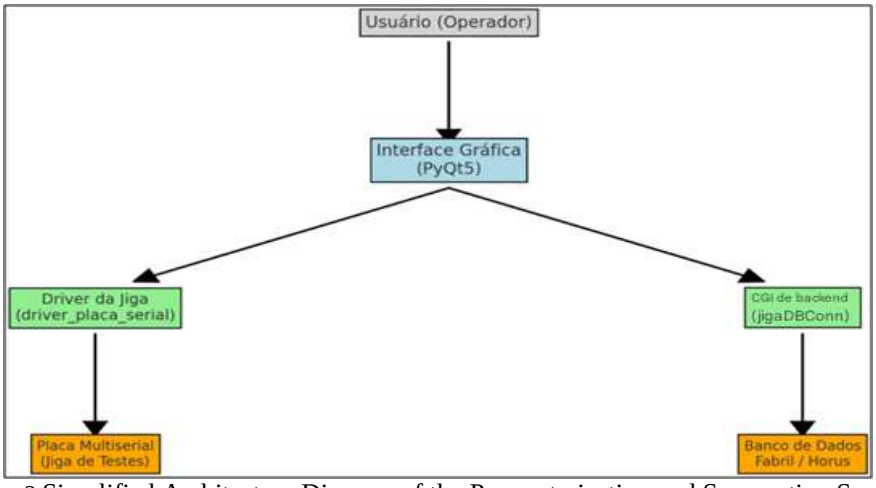


Figure 2 Simplified Architecture Diagram of the Parameterization and Segregation System.
Source: Authors, (2026).

IV.5.5 Main Application – Main

The test results are returned through the carregaMeterStatus slot, which dynamically updates the table with the obtained results, using colors to indicate the success or failure of each meter. At the end of the test batch, if all fields are filled, the interface displays the message "Serial Test Completed!" In the event of critical failures during driver execution, the system calls the erroNoEngine function, which displays an error message to the operator and logs the issue as a critical error. This control reinforces the reliability of the system even under adverse conditions. Another notable feature is the segregar function, which allows the operator to manually record a failed meter outside the automatic flow. When activated, this function opens a segregation window where the operator enters the defective meter's data, and the system records this event in the backend, ensuring full traceability. Additionally, the resetarWidgets function is used at the start of each new work order or in case of a connection error, ensuring the interface is cleared, fields are disabled, and styles are reset for a new work cycle. This structure allows the system to operate in a modular, asynchronous, and visually clear manner. The architecture of paramFechDialog follows industrial software engineering principles, ensuring data consistency, an intuitive operator experience, and seamless integration with other system components. Within this framework, the loginDialog class plays a fundamental role in controlling application access. Its integration with the backend, visual window management, and user feedback mechanisms provide a secure, clear, and efficient interface, aligned with the requirements of industrial environments that handle sensitive data and real-time equipment, as shown in Figures 18 and 19.



Figure 18: Main.
Source: Authors, (2026).

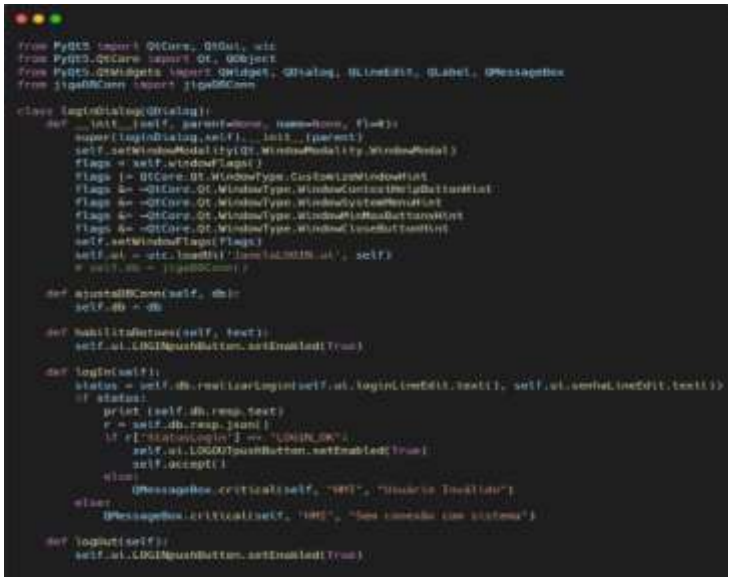


Figure 19: Login.
Source: Authors, (2026).

Therefore, the DialogoSegregarMedidor class represents an important layer of the developed solution. It not only allows operational exceptions to be handled in a user-friendly manner but also reinforces the principles of traceability, data integrity, and compliance with industrial quality and safety best practices. Integrated with the rest of the application through dependency injection and signaling flow, this window helps maintain smooth and secure factory operations even in the face of unexpected events (Figure 20). To ensure consistency and reliability in handling this data, foreign keys (FKs) were applied, and optimized indexes were created based on status, version, and dates, providing better performance for complex queries and greater transactional integrity. Additionally, standardization of possible record states—draft, published, and archived—was established from the beginning, reinforcing the principle of immutability for published versions. This approach, while allowing flexibility during the draft phase, ensured that published versions remained unalterable, guaranteeing the traceability required in a regulated manufacturing environment. To accelerate development and approval cycles, minimal seeds were inserted into the migrations, allowing environments to be initialized with representative data for testing. This strategy reduced the time spent setting up validation scenarios and facilitated the development team's work, enabling them to quickly test functionalities without compromising data integrity. By the end of the first week, the system's relational core was already consolidated, providing a reliable foundation on which subsequent development stages would be built (Figure 21).



Figure 20: Segregate Meter.
Source: Authors, (2026).

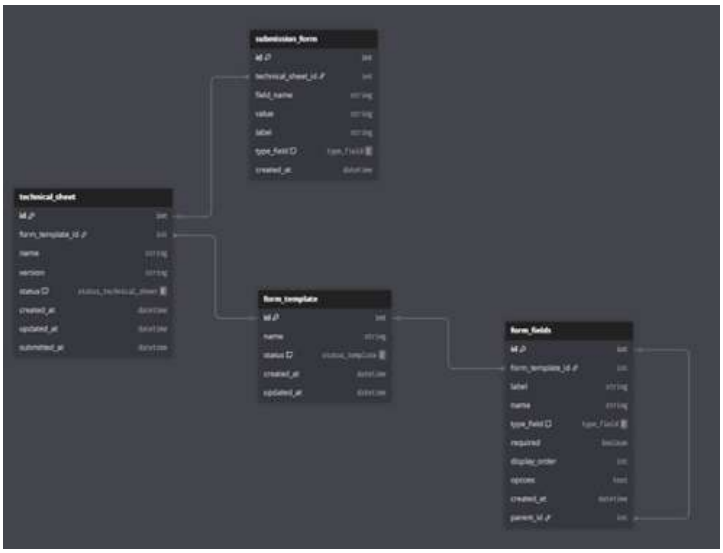


Figure 21: Relational Diagram of the Technical Sheet Module.
Source: Authors, (2026).

IV.6 QUANTITATIVE PERFORMANCE EVALUATION OF THE SYSTEM

Following the gradual implementation of the SIPE modules on the pilot production line for electronic meters, quantitative analyses were conducted to evaluate the system's impact on operational performance. Key performance indicators (KPIs) were monitored before and after the adoption of traceability devices, the Production Control module, and integration with the existing system (SIME). Among the analyzed indicators, the most notable were: productivity rate (meters/hour), average production lead time, rework and rejection rates, average time to identify defective batches, and the unplanned equipment downtime index.

The results indicated a significant reduction in the average lead time of the pilot line, with a decrease in the total time between the start of production and the completion of final testing. An increase in productivity rate was also observed, mainly due to the reduction of unplanned stops and greater predictability in the sequence of operations. In terms of quality, the integration of traceability data and functional testing devices enabled a decrease in rework and rejection rates, as recurring failures were identified more quickly, allowing targeted corrective interventions. Additionally, the time required to locate and isolate problematic batches was significantly reduced, positively impacting the non-conformance analysis process and customer support. Overall, the quantitative results demonstrated that SIPE contributed to increasing the overall efficiency of the pilot line, reducing waste, and enhancing the reliability of production processes, confirming the potential of the proposed approach for industrial environments seeking to adopt Industry 4.0 principles.

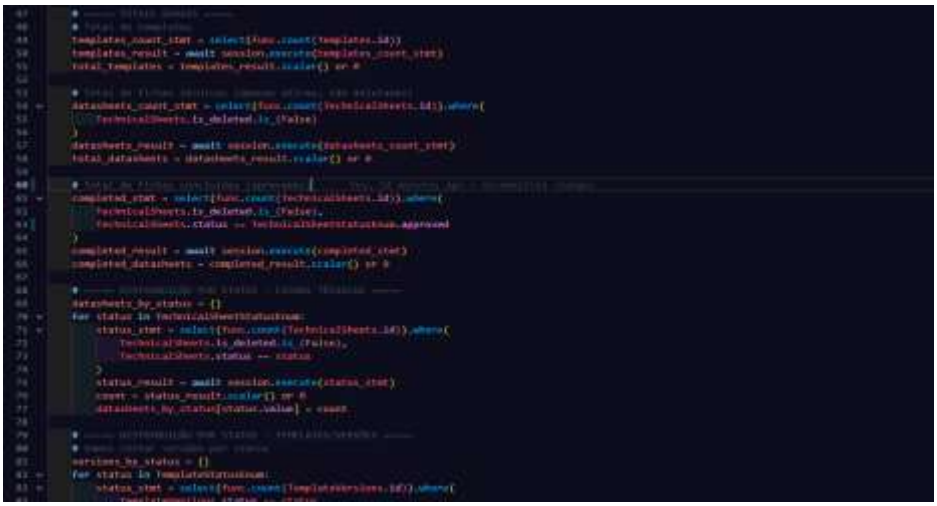


Figure 22: Code for Calculating Statistics and KPIs.
Source: Authors, (2026).

IV.7 IMPLEMENTATION AND EVALUATION OF ARTIFICIAL INTELLIGENCE MODULES

The Artificial Intelligence modules developed within the SIPE framework focused primarily on supporting production planning and replanning, as well as forecasting critical operational variables. Supervised machine learning approaches were used for demand forecasting, cycle time estimation, and identification of production bottleneck patterns. The dataset used to train the models consisted of production order histories, records of meters passing through workstations, machine downtime information, quality test results, and device configuration parameters. These data were extracted from SIME, the new traceability modules, and the Production Control Module, covering a representative period of the pilot line's operation. After data cleaning, normalization, and feature selection, different algorithms were evaluated, including regression, decision trees, and neural network-based models. Model evaluation was performed using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE), considering cross-validation scenarios. The results indicated error levels compatible with practical use in decision-support scenarios, allowing cycle times and workload predictions in advance to support production schedule adjustments. In parallel, classification models were investigated to indicate, based on historical patterns, the likelihood of bottlenecks occurring at specific workstations or an increase in the rejection rate. The integration of the AI modules with the Production Control Module allowed forecasts to be consumed in near real-time by the system, generating automatic replanning suggestions in cases of significant deviations. Although the final decision remained the responsibility of the Production Planning and Control (PPC) team, the use of AI-generated recommendations helped reduce response time to unexpected events and made the planning process more proactive and data-driven.

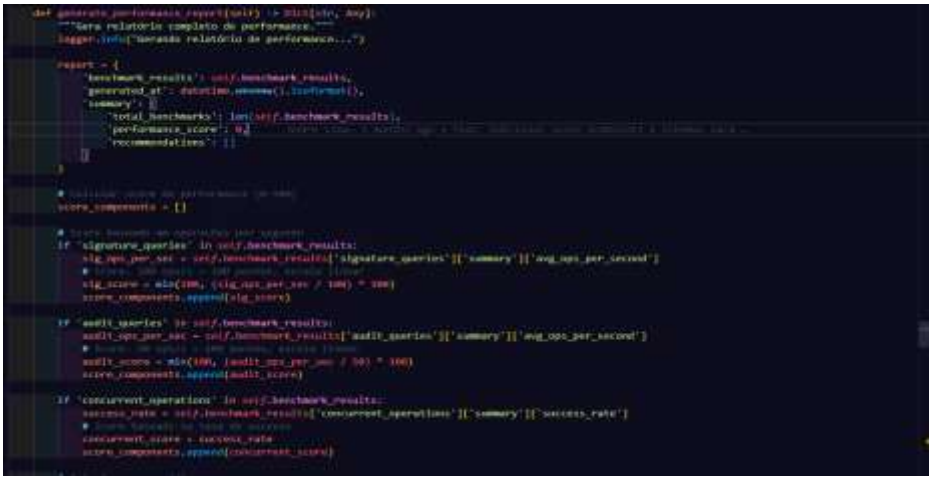


Figure 23: Code for Calculating Performance Score.
Source: Authors, (2026).

IV.8 CASE STUDY ON THE PILOT PRODUCTION LINE OF METERS

To validate the practical applicability of SIPE in a real production environment, a case study was conducted on a pilot production line for electronic meters. The study focused on the integration between the different system modules, demonstrating how the microservices architecture and asynchronous communication between components contribute to operational efficiency and process traceability. The selected pilot line includes workstations with varying levels of automation, from fully connected equipment to devices without native connectivity, which were integrated through the traceability devices developed. SIPE was implemented gradually, starting with the integration of the traceability and production control modules, followed by the implementation of the technical sheet system, and finally the activation of analysis modules and dashboards. One of the most relevant aspects of the case study was the validation of the layered architecture implemented in the system. The presentation layer, developed with Vue and TypeScript, allows operators and managers to view production data in real time through interactive dashboards. The business logic layer, implemented with custom hooks and stores (Zustand), processes and transforms raw data collected from the equipment into actionable information. The communication layer, based on TanStack Query and Axios, manages asynchronous requests to backend services, implementing intelligent caching and automatic retry in case of network failures.

Finally, the backend services layer, composed of the microservices service-auth, service-datasheet, and service-tests, processes business rules and persists data in a distributed manner. The integration between microservices was tested in a real production scenario. When an electronic meter passes through the parametrization device, the system records the event in the traceability module. This event is then processed by the production control module, which updates the production order status and calculates performance indicators. Simultaneously, the technical sheet system is consulted to verify whether the meter specifications comply with the approved template. If any non-conformance is detected, the system generates an alert and records the event in the quality control module, which can trigger automatic production replanning. Service communication was implemented using asynchronous standards with distributed caching (Redis) and exponential backoff retry. For example, when the technical sheet service needs to validate a user profile for approval, it queries the authentication service through the ProfileValidationService. This service implements a cache mechanism with a TTL (Time To Live) of 15 minutes, significantly reducing the number of calls to the external service and improving operation latency. In case of communication failure, the system performs up to 3 attempts with increasing intervals (exponential backoff), ensuring resilience even under unstable network conditions.

During the validation period, a practical scenario demonstrated the effectiveness of the proposed architecture. When a sudden increase in the rejection rate was detected at a functional testing station, SIPE automatically identified the issue through pattern analysis in the traceability data. The production control module correlated the failure events with production batches, identifying that the problem was associated with an improper adjustment on a specific device. The system then issued automatic alerts to the responsible team and suggested partial replanning of production orders, temporarily redirecting the workflow to another set of equipment. This action was executed in less than 15 minutes, from detection to solution implementation, significantly mitigating the impact on daily production targets. The microservices architecture also demonstrated its scalability during the case study. The web service was configured to operate with 4 replicas in the production environment, using Docker Swarm for orchestration and Traefik as a load balancer. This setup allowed the system to maintain high availability even during demand peaks, automatically distributing the load among instances. The technical sheet service, in turn, was dimensioned to process up to 50 signing operations per second, as validated by performance tests, ensuring that the system would not become a bottleneck during critical approval operations. Beyond quantitative benefits, the case study highlighted qualitative gains related to process visibility, standardization of records, and support for decision-making.

Operators and engineers gained an integrated view of the production line, with real-time access to each equipment's data, test results, and production order status. This visibility facilitated root cause identification of recurring problems and allowed prioritization of continuous improvement actions based on objective data. The case study also validated the effectiveness of the electronic signature system integrated with SIPE. During the validation period, over 500 template and technical sheet approvals were processed, all with full traceability through audit logs. The SHA-256 hash system ensured document integrity, and automatic invalidation of signatures when templates were modified ensured that only approved versions were used on the production line. This functionality was particularly relevant to ensure compliance with regulatory requirements and industry quality standards. In summary, the case study on the pilot line demonstrated that the microservices architecture, combined with asynchronous communication, distributed caching, and real-time processing, provides a robust, scalable system capable of supporting the demands of an industrial production environment. The integration of SIPE's different modules enabled more efficient production management, with increased visibility, traceability, and responsiveness to unexpected events, confirming the potential of the proposed approach for digital transformation in manufacturing environments.

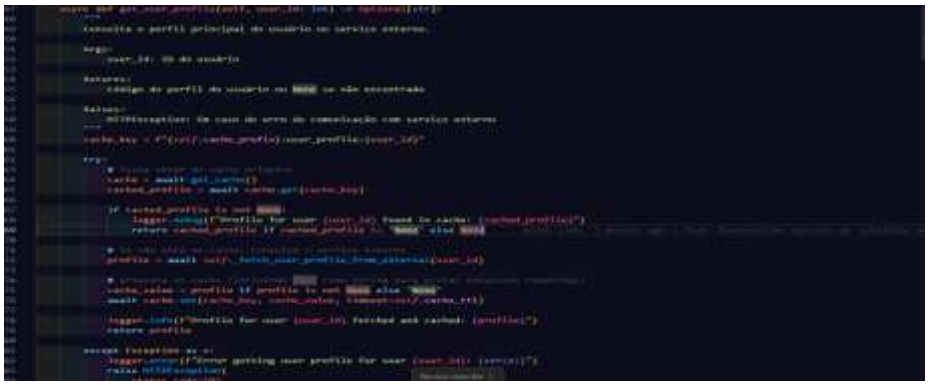


Figure 24: Integration code between microservices with cache and retry.
Source: Authors, (2026).

V. CONCLUSÕES

It is concluded that the development of an Automated System for Control and Production Planning of Electronic Meters with Artificial Intelligence represents a significant advancement in the modernization of industrial processes, aligning technology, efficiency, and reliability. The adoption of the Internet of Things (IoT) enabled the interconnection of production equipment, allowing real-time monitoring, continuous information exchange, and decision-making based on accurate and up-to-date data. The use of traceability and parametrization devices, combined with the asynchronous PIMA serial output, contributed to rigorous control of the production flow, ensuring the individual identification of each electronic meter throughout all stages of the process. This was essential for strengthening traceability, providing a complete history of production, configuration, and testing, as well as supporting quality control through early fault detection and reduced rework.

The integration of electronic meters with the Raspberry Pi 5 (8GB) proved to be an efficient and low-cost solution, offering high performance for data collection, processing, and analysis. The communication interface, together with integration via industrial protocols, ensured interoperability between heterogeneous systems, facilitating automation and control of the production process. Through data collection and processing, combined with artificial intelligence techniques, it was possible to enhance production planning, optimize resources, and increase the reliability of management information. Connectivity with APIs, networks, and digital services expanded system scalability, allowing integration with corporate platforms and cloud environments. Thus, the proposed system proves to be robust, flexible, and aligned with Industry 4.0 concepts, promoting greater operational efficiency, final product quality, and competitiveness in the electronic meter sector.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Jandecy Cabral Leite, Marcelo Maia do Nascimento, Alan Ferreira Pinheiro Tavares, André Barbosa de Lima, Matheus Pereira de Souza, Daniel Gomes Gulart and Marcos Castro de Menezes.
Methodology: Alan Ferreira Pinheiro Tavares, André Barbosa de Lima, Matheus Pereira de Souza.
Investigation: Jandecy Cabral Leite, Marcelo Maia do Nascimento, Alan Ferreira Pinheiro Tavares, André Barbosa de Lima, Matheus Pereira de Souza, Daniel Gomes Gulart and Marcos Castro de Menezes.
Discussion of results: Jandecy Cabral Leite, Marcelo Maia do Nascimento, Alan Ferreira Pinheiro Tavares.
Writing – Jandecy Cabral Leite.
Writing – Review and Editing: Jandecy Cabral Leite and Marcelo Maia do Nascimento.
Resources: Jandecy Cabral Leite and Marcelo Maia do Nascimento.
Supervision: Jandecy Cabral Leite, Marcelo Maia do Nascimento and Alan Ferreira Pinheiro Tavares.
Approval of the final text: Jandecy Cabral Leite, Marcelo Maia do Nascimento.

VII. AGRADECIMENTOS

To ITEGAM and Wasion da Amazônia Electronic Instruments Industry LTDA for the financial support of the R&D&I Project through the Information Technology Law (Law No. 8,248/1991), which encourages investment in research, development, and innovation (R&D&I) in the information and communication technology (ICT) sector through tax incentives and financial credits for companies in the Manaus Industrial Pole (PIM).

VIII. REFERENCES

[1] Saldivar, A. A. F., Li, Y., Chen, W. N., Zhan, Z. H., Zhang, J., & Chen, L. Y. (2015, September). Industry 4.0 with cyber-physical integration: A design and manufacture perspective. In *2015 21st international conference on automation and computing (ICAC)* (pp. 1-6). IEEE.

[2] Zhang, K., Shi, Y., Karnouskos, S., Sauter, T., Fang, H., & Colombo, A. W. (2022). Advancements in industrial cyber-physical systems: An overview and perspectives. *IEEE Transactions on Industrial Informatics*, *19*(1), 716-729.

[3] Leitão, P., Colombo, A. W., & Karnouskos, S. (2016). Industrial automation based on cyber-physical systems technologies: Prototype implementations and challenges. *Computers in industry*, *81*, 11-25.

[4] Shadravan, A., & Parsaei, H. R. (2023). The Paradigm Shift From Industry 4.0 Implementation to Industry 5.0. *Application of Emerging Technologies*, *115*, 1.

[5] Nagy, M. (2025). Industria mobilis: Integrating cyber-physical systems, data fusion, and autonomous robotics in Slovak Industry 4.0. *Economics, Management and Sustainability*, *10*(2), 26-46.

[6] Alvarez, M. L., Sarachaga, I., Burgos, A., Estévez, E., & Marcos, M. (2016). A methodological approach to model-driven design and development of automation systems. *IEEE Transactions on Automation Science and Engineering*, *15*(1), 67-79.

[7] Alvarez, M. L., Estévez, E., Sarachaga, I., Burgos, A., & Marcos, M. (2013). A novel approach for supporting the development cycle of automation systems. *The International Journal of Advanced Manufacturing Technology*, *68*(1), 711-725.

[8] Melo, G., Nascimento, N., Alencar, P., & Cowan, D. (2023). Identifying factors that impact levels of automation in autonomous systems. *IEEE Access*, *11*, 56437-56452.

[9] Morrar, R., Arman, H., & Mousa, S. (2017). The fourth industrial revolution (Industry 4.0): A social innovation perspective. *Technology innovation management review*, *7*(11), 12-20.

[10] Philbeck, T., & Davis, N. (2018). The fourth industrial revolution. *Journal of International Affairs*, *72*(1), 17-22.

[11] Castelo-Branco, I., Oliveira, T., Simões-Coelho, P., Portugal, J., & Filipe, I. (2022). Measuring the fourth industrial revolution through the Industry 4.0 lens: The relevance of resources, capabilities and the value chain. *Computers in industry*, *138*, 103639.

[12] Mourtzis, D., Vlachou, E., & Milas, N. J. P. C. (2016). Industrial big data as a result of IoT adoption in manufacturing. *Procedia cirp*, *55*, 290-295.

[13] Ur Rehman, M. H., Yaqoob, I., Salah, K., Imran, M., Jayaraman, P. P., & Perera, C. (2019). The role of big data analytics in industrial Internet of Things. *Future Generation Computer Systems*, *99*, 247-259.

[14] Khan, M., Wu, X., Xu, X., & Dou, W. (2017, May). Big data challenges and opportunities in the hype of Industry 4.0. In *2017 IEEE International Conference on Communications (ICC)* (pp. 1-6). IEEE.

[15] Malek, Y. N., Kharbouch, A., El Khoukhi, H., Bakhouya, M., De Florio, V., El Ouadghiri, D., ... & Blondia, C. (2017). On the use of IoT and big data technologies for real-time monitoring and data processing. *Procedia computer science*, *113*, 429-434.

[16] Chen, Y. (2020). IoT, cloud, big data and AI in interdisciplinary domains. *Simulation Modelling Practice and Theory*, *102*, 102070.

[17] Lampropoulos, G. (2023). Artificial intelligence, big data, and machine learning in industry 4.0. In *Encyclopedia of data science and machine learning* (pp. 2101-2109). IGI Global.

[18] Kibria, M. G., Nguyen, K., Villardi, G. P., Zhao, O., Ishizu, K., & Kojima, F. (2018). Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks. *IEEE access*, *6*, 32328-32338.

[19] Venkatesh, D. A. N. (2019). Reimagining the future of healthcare industry through internet of medical things (IoMT), artificial intelligence (AI), machine learning (ML), big data, mobile apps and advanced sensors. Artificial Intelligence (AI), Machine Learning (ML), Big Data, Mobile Apps and Advanced Sensors (October 28, 2019).

[20] Manoharan, K. G., Nehru, J. A., & Balasubramanian, S. (2021). Artificial intelligence and IoT. *Springer*, *10*, 978-981.

[21] Rane, N. (2023). Integrating leading-edge artificial intelligence (AI), internet of things (IOT), and big data technologies for smart and sustainable architecture, engineering and construction (AEC) industry: Challenges and future directions. *Engineering and Construction (AEC) Industry: Challenges and Future Directions (September 24, 2023)*.

[22] Lăzăroiu, G., Andronie, M., Iatagan, M., Geamănu, M., Ștefănescu, R., & Dijmărescu, I. (2022). Deep learning-assisted smart process planning, robotic wireless sensor networks, and geospatial big data management algorithms in the internet of manufacturing things. *ISPRS International Journal of Geo-Information*, *11*(5), 277.

[23] Usuga Cadavid, J. P., Lamouri, S., Grabot, B., Pellerin, R., & Fortin, A. (2020). Machine learning applied in production planning and control: a state-of-the-art in the era of industry 4.0. *Journal of Intelligent Manufacturing*, *31*(6), 1531-1558.

[24] Cioffi, R., Travaglioni, M., Piscitelli, G., Petrillo, A., & De Felice, F. (2020). Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions. *Sustainability*, *12*(2), 492.

[25] Shi, H., Cao, G., Ma, G., Duan, J., Bai, J., & Meng, X. (2022). New progress in artificial intelligence algorithm research based on big data processing of IOT systems on intelligent production lines. *Computational intelligence and neuroscience*, *2022*(1), 3283165.

[26] Zhao, X., Wang, R., Wang, Y., & Wang, L. (2020, September). Design and realization of metering production plan management system. In *Journal of Physics: Conference Series* (Vol. 1633, No. 1, p. 012155). IOP Publishing.

[27] Vesikkala, T. (2017). Lean Optimization for Electric Meter Production.

[28] O'driscoll, E., & O'donnell, G. E. (2013). Industrial power and energy metering–a state-of-the-art review. *Journal of cleaner production*, *41*, 53-64.

[29] Duncan, B. K., & Bailey, B. G. (2004). Protection, metering, monitoring, and control of medium-voltage power systems. *IEEE Transactions on Industry Applications*, *40*(1), 33-40.

[30] Benyoucef, D., Bier, T., & Klein, P. (2010, June). Planning of energy production and management of energy resources with Smart Meters. In *2010 International Conference on Advances in Energy Engineering* (pp. 170-173). IEEE.

[31] Li, D., Qu, J., Su, Q., Du, B., Guo, L., Zhang, Q., & Cai, W. (2018, July). Research on Smart Electric Energy Meters Manufacturing Technology Program Based on RFID. In *2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC)* (pp. 1538-1541). IEEE.

[32] Oteri, O. J., Onukwulu, E. C., Igwe, A. N., Ewim, C. M., Ibeh, A. I., & Sobowale, A. (2023). Cost optimization in logistics product management: strategies for operational efficiency and profitability. *International Journal of Business and Management. Forthcoming*.

[33] Gamal, H., Albahri, S., Essa, H., Srivastava, M., Alghamdi, N., & Shaikh, W. (2024, February). Driving operational excellence: Cost savings and performance optimization through tripping and connection time optimization. In *International Petroleum Technology Conference* (p. D021S077R002). IPTC.

[34] Kaydos, W. (2020). *Operational performance measurement: increasing total productivity*. CRC press.

[35] Liao, H. C., Lim, Z. Y., Hu, Y. X., & Tseng, H. W. (2018, July). Guidelines of automated optical inspection (AOI) system development. In *2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP)* (pp. 362-366). IEEE.

[36] Abd Al Rahman, M., & Mousavi, A. (2020). A review and analysis of automatic optical inspection and quality monitoring methods in electronics industry. *Ieee Access*, *8*, 183192-183271.

[37] Ronchi, M., Regattieri, A., Gamberi, M., & Cafarella, C. (2025). Multi-criteria selection of image acquisition technologies for automatic optical inspection in assembly lines. *The International Journal of Advanced Manufacturing Technology*, 1-20.

[38] Goti, A. (2025). Automated Optical Inspection (AOI) Based on IPC Standards.

[39] Oluyisola, O. E., Bhalla, S., Sgarbossa, F., & Strandhagen, J. O. (2022). Designing and developing smart production planning and control systems in the industry 4.0 era: a methodology and case study. *Journal of Intelligent Manufacturing*, *33*(1), 311-332.

[40] Alaji, F. (2024). *Enhancing Energy Efficiency in Production Lines Through Advanced Process Monitoring and Control System* (Doctoral dissertation, Magyar Agrár- és Élettudományi Egyetem).

[41] Bueno, A., Filho, M. G., Carvalho, J. V., & Callefi, M. (2021). Smart production planning and control model. In *Perspectives and Trends in Education and Technology: Selected Papers from ICITED 2021* (pp. 253-267). Singapore: Springer Singapore.

[42] Hoffmann, M. (2019). *Smart agents for the Industry 4.0* (pp. 226-297). Springer: Berlin, Germany.

[43] Lai, C. A. (2022). System level of intelligent manufacturing. In *Intelligent Manufacturing* (pp. 53-105). Singapore: Springer Nature Singapore.

[44] Sundar, P. S., Chowdhury, C., & Kamarthi, S. (2024). Industrial Internet of Things Enabled Kata Methodology of Assembly Line Productivity Improvement: Insights from a Case Study. *Processes*, *12*(11), 2611.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 519-528. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3166

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

QUANTIFYING THE CREDIBILITY OF E-LEARNING SYSTEMS USING THE BERT MODEL

Piyush Singh[1], Rohan Pachisiya[2], Sagar Nehra[3] and Vibha Gaur*[4]

[1, 2, 3, 4]Department of Computer Science, Acharya Narendra Dev College, University of Delhi, New Delhi-110019, India.

[1]https://orcid.org/0009-0002-0769-8272, [2]https://orcid.org/0009-0000-6209-1866
[3]https://orcid.org/0009-0002-7907-6755, [4]https://orcid.org/0000-0001-6668-9339

Email: piyush.ae-1237@andc.du.ac.in, rohan.ae-1242@andc.du.ac.in, sagar.ae-1244@andc.du.ac.in, *vibhagaur@andc.du.ac.in

ARTICLE INFO

Article History
Received: December 18, 2025
Revised: December 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
User autonomy,
Dark pattern,
BERT classifier,
Educational websites,
Credibility index (CI).

ABSTRACT

Trust has become crucial in a digitized world for the sustainability of online platforms. As the web interfaces have been playing a vital role in day-to-day activities, ethical design has become essential to protect user autonomy and promote informed decision-making by them. Dark patterns are deceptive design strategies that can harm users' trust and are dangerous, especially in the e-learning environment. This paper presents a refined Bidirectional Encoder Representations of Transformers (BERT) classifier to automatically detect dark patterns on educational online systems. The framework starts with web scraping of the digital interface followed by organized preprocessing, such as content extraction, text cleansing, normalization, and tokenization. An algorithm to calculate a Credibility Index (CI) of e-learning systems is proposed based on the frequency and the severity of perceived dark patterns. The online systems are categorized into one of the three threat levels—*Safe*, *Moderate*, or C*ritical*, which gives clear indication to users regarding the trustworthiness of the site. Using the proposed framework, a customized educational website called SkillNest was developed to predict user trust. It was classified as *Moderate* due to its CI value of 0.64. This work may help developers in enhancing transparency and trust in educational technology by reducing the manipulative practices for developing e-learning systems.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Dark patterns are untrustworthy and manipulative user interface (UI) design methods that are used by digital platforms to control user behavior in a way that can harm their choices, threaten trust, and break down faith [1-3]. The term dark pattern was given by Brignull [4] to describe ways of using mind tricks and psychological rules to make the platform earn the most money, usually at the cost of what's best for the users [2]. They include biases caused by social pressure, as well as the way choices are presented to push users toward performing things they wouldn't normally do [5], [6]. Also, such manipulative actions increase digital inequality, affecting digitally illiterate groups, and increasing the need for stricter rules on how platforms need to be designed [7]. Dark patterns restrict the users' rights to make a fair decision. Difficult procedures make it difficult for users to leave online services and thus limit their control over their personal data and online activities [6], [8].

These malpractices raise serious concerns about designing guidelines for developing online systems [2], [7]. Even the Department of Consumer Affairs has banned 13 different types of dark patterns. Since many people are not aware of these malpractices, there is a strong demand for authentic websites. Though various studies in literature report the types and effects of dark patterns [9], their automated identification in real time and measuring their impact has remained a challenge [3]. Manual detection of dark patterns requires a lot of time and effort and may be error-prone for complicated digital interfaces [10]. To address these issues, designers and customers need automated tools to detect and fix the causes of dark patterns [1]. Without these tools, people may find it challenging to follow the law and make informed decisions due to the increasing complexity of these dishonest strategies [11].

Finding the UI design malpractices on a huge scale is now feasible due to improvements in tools that understand natural language and learn from data, known as machine learning (ML). Large Language Models (LLMs), such as BERT [12], are capable of interpreting

the meaning and context of language [13]. This feature may be utilised in discovering elusive context-based dark patterns, which are otherwise difficult to catch by simple rule-based systems [14]. Also, these language models may offer deeper insights into these patterns of user behavior. These transformer models have also been utilised in other applications. For example, the BERT model has been used to extract and categorize large amounts of critical information from social media platforms [15]. Research by Kaur et al. [16] has showcased the application of the BERT model for quantifying the quality of online learning systems.

The proposed study presents an automated system to detect dark patterns and measure their impact to classify the educational platforms. It involves gathering of data, followed by preprocessing, including cleaning up the text and standardizing it using normal methods to make sure the data used is of high quality [17]. The refined BERT model calculates the chance that the user interface text is manipulative [18]. In addition, the proposed system also calculates the value of CI for digital platforms. The CI is a trust score that indicates the trustworthiness of a website by looking at the frequency and seriousness of the dark patterns. The digital platforms are assigned a threat level—*Safe*, *Moderate*, or *Critical*, based on the value of CI. It facilitates users' understanding of whether they can safely access the site or need to be careful or avoid the site altogether.

This work focuses on building credibility and trust by bringing it to the education world, where manipulative design practices can harm a student's confidence and engagement in e-learning platforms. Introducing the CI may help the students in using the trustworthy websites only. Though there are numerous ways of building trust. Sharma et al. [19] proposed that faith and trust in e-commerce can be developed by fixing problems like missing information and addressing bad reputations on business websites. Though this study has focused on educational websites, the framework is flexible and can be used for any digital platform. The main aim of the study is to develop an automatic system for predicting the credibility of educational websites. This research mainly addresses the problems caused by poor design on the educational platforms by focusing on the following objectives:

1. To build a system that automatically detects the dark patterns on learning websites using smart language models (like BERT).
2. To compute the CI of the online system based on the frequency and severity of the dark patterns identified.
3. To classify the online e-learning systems into three levels—*Safe*, *Moderate*, or *Critical* based on the value of CI.
4. To support fairness and transparency in digital educational environments.

The paper is divided into several parts: Section 2 provides a literature review of past research work on dark patterns and methods for finding them automatically. Section 3 presents the methodology for predicting the credibility of the educational online systems, while Section 4 discusses the test results. Finally, Section 5 concludes the paper with a summary of its main contributions.

II. LITERATURE REVIEW

Dark patterns are malign strategies that affect user trust. A lot of research is being done to identify and categorize the dark patterns. However, there are still gaps in terms of applying effective technology to identify these patterns in particular contexts. Luguri and Strahilevitz [2] were the first to recognize and categorize dishonest website layouts, such as the "sneak into basket," which produced the fundamental concepts for future studies. This classification was later expanded by Narayanan et al. [1] through interviews with experts, revealing 12 ethical issues, among which were “compulsory continuity” and “interface interference.” Mathur et al. [20] provided comprehensive evidence for this by analysing 11,000 shopping websites and arriving at 15 categories of dark patterns. Their results showed the effect of manipulative user interfaces on cognitive biases, thus emphasizing the immediate need for fast, scalable, and automated detection systems.

Gunawan et al. [21] presented that the dark patterns are significantly different between conventional sites and mobile applications, which means that lying tactics differ based on the platform. These studies provided valuable background knowledge towards the definition and description of dark patterns but failed to present an implementable, machine learning-powered, quantifiable assessment system in real time. It is in this regard that the necessity to have an automated system that constantly identifies and analyses dark patterns on digital platforms arises. Research in the field of behavioural economics has deeply searched for the psychological reasons of why dark patterns are used. A. Golandaz and U. Sharma [9] showed the way the dark patterns capitalize on cognitive biases such as loss aversion and default effects, and Greenberg et al. [8] have pointed to their usage in spatial interfaces in terms of close interactions. Mobile app analysis by Brignull [4] discovered that 95% of apps are designed in a manipulative fashion.

The work by Nouwens et al. [22] focused on the study of GDPR consent pop-ups using web scraping to reveal huge fraudulent cases. All these studies demonstrate how dark patterns restrict user autonomy, but studies continue to be limited by the differences in susceptibility by demographic, especially in educational contexts. Current methods of detection have numerous issues related to scalability and are not effective in terms of extensive tracking. Newer machine learning concepts have prospects and need to be explored fully as suggested by Mathur et al. [20], but have limitations due to the existing data sets. The proposed work bridges this gap by using the language model, which is sophisticated. Additional related literature in misinformation detection by Soni et al. [23] depicted that BERT based models and sentiment characterization can achieve good performance in identifying and interpreting manipulative information on social media using natural language processing (NLP) based approaches in detecting trust-violating patterns.

Ready-to-train transformer architectures, such as BERT [13] and RoBERTa presented by Liu et al. [14], have greatly contributed to the understanding of language formations, especially where manipulative design hints are hidden in minor textual details. These models generate deep contextual settings and support specific classification adaptation, as given by Sun et al. [24]. The introduction of the ELECTRA model by Clark et al. [25] enhances training effectiveness using substituted token detection. Though generally used for NLP, such models have not been well studied in identifying dark patterns. In this paper, these architectures are being used for autonomous identification of fraudulent user interface designs. Trust and credibility are important to the online platform, particularly in the face of fake features of designs, such as dark patterns, which possess more power to impact user behavior.

Little and Green [26] presented a framework for trust that is based on concepts like faith and leadership that doesn't depend on formal titles, reliability, and being seen as an expert. This framework shows that the developers have an opportunity to promote informed decision-making via reliable design. P. Paithane [27] stressed the importance of a structure-based framework that relies on both interaction

and behavioural data to measure faith in the online setting. Most critically, Corritore et al. [28] came up with a scale that can be used to measure customer trust on websites in accordance with the fact that a reliable tool is lacking in gauging online trust. Bedi and Gaur [18] extended the knowledge of trust among agent systems by including dimensions such as usability, performance, integrity, and security. It is on these grounds that the need to incorporate trust modelling into the process of identifying manipulative interfaces is amplified.

Lanier [29] suggested user-focused platforms that give importance to transparency and independence, while T. Nyström and A. Stibe [7] highlighted the wider risks of surveillance-based design with their ethical assessment. The Department of Consumer Affairs of the Government of India has also provided official guidelines in 2023 prohibiting certain dark patterns. As noted by S. Han and C. Anderson [30] regarding methodological biases in digital research, technical challenges with data collection persist. Scalable and integrated automated measurement tools and trust appraisal are still minimal, especially in education, where fairness is of great importance. Building upon these insights, this study examines an integrated approach, which is a combination of transformer-based classification and a CI intended to detect, measure, and report manipulative design practices, which is crucial for strong assessment of trust in e-spaces.

III. MATERIALS AND METHODS

Detection of dark patterns is crucial for the sustainability of the digital platforms. As the manual methods for identification of dark patterns are neither effective nor scalable. This research proposes an automated framework for dark pattern detection using the BERT model. The first step of the proposed work involves the extraction of the relevant e-content followed by preprocessing to clean and standardize the text. An optimally adjusted BERT-based model is utilised to identify instances of dark patterns with accuracy. To assess the general reliability of the online platforms, the Credibility Index algorithm is presented that considers the frequency and the severity of the discovered patterns. Figure 1 demonstrates the suggested framework to predict the credibility of online platforms. The framework comprises of the following steps: Data Extraction, Preprocessing of data, Pre-Trained BERT Model, Fine-Tuning of Model, Segregation of Dark Patterns, and Computation of CI, which are explained below.

III.1 DATA EXTRACTION

Data extraction involves downloading the raw HTML content of the URL on the online platform with the help of web scraping software like BeautifulSoup and Scrapy. This may consist of descriptive information, user reviews and policy documents. The raw HTML content is then changed to structured form.

III.2 PREPROCESSING OF DATA

The extracted data is processed using preprocessing operations to make it suitable for machine learning tasks. The raw content is first cleaned, i.e. HTML tags, special characters and other irrelevant content such as advertisements or navigation menus are removed. Linguistic noise is eliminated by normalizing the text, turning all letters to lowercase, and eliminating stop words. Lastly, the text is divided into a set of separate words or phrases called tokens, which can be interpreted by the language model.

III.3 PRE-TRAINED BERT MODEL

BERT is a language model that provides the semantic meaning of text by looking at the preceding and succeeding words simultaneously. It is a two-way text-processing model that processes the text in both directions at once and enables to see the whole picture of each token. It has been trained on large-scale corpora, such as BookCorpus and English Wikipedia, to predict and identify relationships between sentences. In this paper, the BERT-base-uncased model is used because of its good performance in natural language understa nding and high transfer learning, making it appropriate for detection of the dark patterns. The extracted data is initially processed with cleaning and tokenizing operations and then sorted into a prediction set for fine-tuning the model.

III.4 FINE-TUNING BERT MODEL

Even though the pre-trained BERT model provides an outstanding initial point for understanding language in general, it needs to be fine-tuned for detecting and classifying dark patterns. Fine-tuning involves adjusting the learning parameters of a model based on domain-specific knowledge so that it can learn to pick up the finer linguistic details and contextual differences that may signal deceptive interface designs.

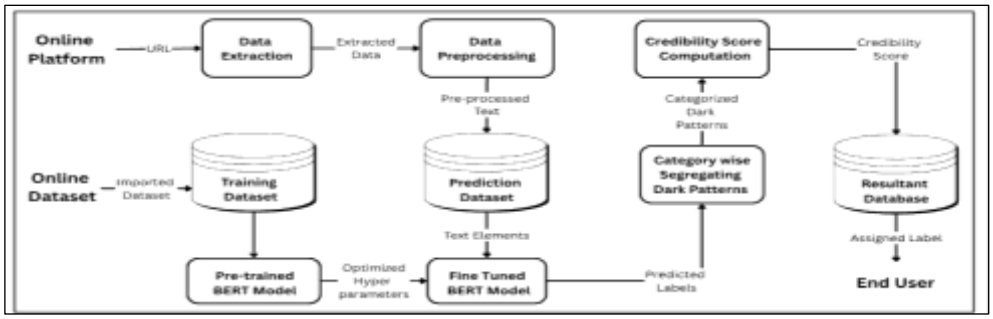


Figure 1: Proposed framework for determining the credibility of an online system.
Source: Authors, (2026).

At this step, the model does not change the transformer encoder architecture but instead changes the output layer by a task-specific classification head that is adapted to binary classification. The annotated dataset is trained with the optimal set of hyperparameters

so that the model can change the representations it learns based on the language and contextual clues that are commonly used in deceptive design.

III.5 SEGREGATION OF DARK PATTERNS

After identifying dark patterns with the help of the fine-tuned BERT model, these are classified according to the predefined taxonomies. This step is important in going beyond identification to find out the cause of the manipulative design. A labelled classification schema is used to map each instance to one of a set of predefined categories that enables the analysis without sacrificing the context of each manipulation strategy. This taxonomy helps to evaluate the behavioural capabilities of the interface of websites not only by their occurrence but specifically by the types of dark patterns used.

III.6 COMPUTATION OF CREDIBILITY INDEX (CI)

CI is a composite measure that shows the overall interface credibility of a platform based on the occurrence, frequency, and nature of dark patterns. It is calculated through the aggregation of weighted scores of identified dark patterns. Each dark pattern has been assigned a weight, which is based on empirical analysis and corresponds to the extent of loss the given dark pattern can bring. Weights may be procured from the domain experts. The proposed method uses the frequency (F) of each dark pattern and the weight of the dark pattern to derive a Severity Score. The values of the severity score and CI are computed using Eq. (1) and (2), respectively.

$$\text{Severity } Score = \left\{ \sum_{i=1}^{\lambda} (F_i \times W_i) \right\} \tag{1}$$

$$Credibility\ Index = 1 - Severity\ Score \tag{2}$$

Where:

- F_i – Frequency of the i^{th} dark pattern
- W_i – Weight of the i^{th} dark pattern
- λ – Type of unique dark patterns

Algorithm 1: Computation of credibility index.

Let Fi be the frequency of the i^{th} dark pattern, λ represent the maximum possible number of dark pattern types detectable by the system, and let W_i be the weight of the i^{th} dark pattern on the webpage.

for i = 1 to λ do

*SeverityScore ← SeverityScore + (Fi * W_i)*

end

CredibilityIndex ← 1 – SeverityScore

Return CredibilityIndex ∈ [0, 1]

Source: Authors, (2026).

Essentially, the Severity Score measures the disbelief on a digital platform in terms of the footprint of dark patterns, while CI is the extent to which the interface of a platform is trustworthy. The pseudocode of appraising the CI on the basis of the occurrence and severity of the dark patterns identified is given in Algorithm 1. The final value of CI lies between 0 and 1, which indicates the credibility of an online platform. More value of CI indicates a highly trustworthy site and vice-versa. It helps users to make wise choices and promotes the establishment of more equitable and transparent digital landscapes. A threat level is assigned to each website using CI to support the interpretation. CI is computed for each of the webpages of the site, and an average is taken to assign an overall threat level to the platform. Three levels of threat—*Safe*, *Moderate* and *Critical* are assigned, with *Critical* being the most manipulative design. The classification enables users to be aware of potential risks associated with the online systems.

IV. IMPLEMENTATION AND RESULTS

To see the application of the proposed framework, the bespoke website *SkillNest*, consisting of three webpages with a variety of dark patterns to allow more comprehensive and realistic evaluation, was created using Python at the backend for extracting and processing data and running algorithms. The user interface was designed using JavaScript, HTML, and Flask. And MySQL was used as the structured data storage, which guarantees the efficiency and scalability. The proposed framework aimed at categorizing seven classes of dark patterns, which are chosen due to their significance and frequency in digital user interfaces [20]. These are discussed in brief.

1. *Sneaking*: The deliberate concealment or omission of critical information to mislead users, such as hiding additional charges until the final checkout stage.
2. *Urgency*: It induces fake time constraints and forces them to take impulsive decisions.
3. *Misdirection*: Visual or textual indicators are provided to distract their attention away from relevant content, often leading them to take inadvertent actions.

4. *Social Proof*: It exploits users' psychological nature to follow others in unseen or unknown situations by showing fake or exaggerated data influencing the users' decisions.
5. *Scarcity*: It persuades the users by displaying limited stock available regardless of the actual availability.
6. *Obstruction*: It creates impediments to user workflows, thus intentionally obscuring crucial actions like subscription cancellations or privacy setting modifications.
7. *Forced Action*: It compels users to perform tasks such as mandatory account creation before accessing content or services. that are not directly relevant for accessing the primary intent,

These design patterns are often used to influence the user choices in a subtle or coercive manner. Each category of dark pattern targets a particular form of deception and could be beneficial for interface-based manipulation. Such a categorization schema is essential for model training, as well as credibility computation. The framework allows analysing manipulative design strategies in a granular and consistent manner by mapping identified patterns to particular behavioural strategies and improving the intelligibility and the responsibility of the detection system. The next step after the classification is the technical implementation of the framework with a strong system architecture. The proposed system used the following architecture:

1. *Content Extraction & Preprocessing*: Raw HTML content of the website SkillNest is scraped, parsed, and cleaned to retain relevant text and semantic elements.
2. *Model Integration*: A fine-tuned BERT model identifies and classifies textual elements indicative of dark patterns.
3. *Database Storage*: Model predictions, metadata, and user feedback are stored in a MySQL database.
4. *Interface Layer*: The front end is a Flask-based application that links the users to the backend services through the use of the HTTP paths so that users interact seamlessly with the real-time outcomes.

An uncased BERT model was fine-tuned utilising TensorFlow 2.0 plus the Transformers library in a Python environment and was used for the classification process. A snapshot of the implementation is presented in Figure 2.

```
train_InputExamples, validation_InputExamples = convert_data_to_examples(
    train, test, DATA_COLUMN, LABEL_COLUMN
)

train_data = convert_examples_to_tf_dataset(list(train_InputExamples), tokenizer, max_len=128)
train_data = train_data.shuffle(100).batch(16)

validation_data = convert_examples_to_tf_dataset(list(validation_InputExamples), tokenizer, max_len=128)
validation_data = validation_data.batch(16)

model.compile(optimizer=tf.keras.optimizers.experimental.AdamW(learning_rate=3e-5, weight_decay=0.01),
              loss=tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True),
              metrics=['accuracy'])

model.fit(train_data, epochs=3, validation_data=validation_data)
```

Figure 2: The BERT classifier configuration parameters.
Source: Authors, (2026).

A labelled dataset [20] was utilised from the publicly available Yamana Lab repository [31] to fine-tune the BERT model. The resulting data set consisted of a combination of manipulative and non-manipulative elements of UI text, each of which was labelled as either dark (1) or not dark (0) by domain experts having knowledge of interface design and regulatory standards. Some of the examples of dark patterns found in the dataset were sneaking, urgency messages, social proof, forced action, and obstruction, and the non-dark patterns included transparent consent prompts and user-respecting design choices. To confirm the practical relevance of the trained model, an operating assessment pipeline was created in Python. This included the process of loading the fine-tuned classifier, preprocessing text found within HTML, and outputting predictions to find out the dark patterns. The 5-fold cross-validation was used to train the classifier so that it can generalize and minimize overfitting.

Every fold was trained on part of the data, and the remaining data was used for validation. This trial process was used to achieve the stable performance and emphasized the capabilities of the model to recognize obscure manipulative signals and context-sensitive wording that are employed in dark patterns. Table 1 provides the optimal hyper-parameter configuration that was used to fine-tune the BERT classifier, with the decisions being made based on performance and efficiency. The length of the maximum sequence was fixed to 128, which is the highest number of tokens that were handled each time a single input is given. *AdamW* optimizer is a weight decay adaptive optimization algorithm that was used in updating parameters. The classification objective function—*Cross-Entropy Loss* was used. The training was done in three epochs, and the architecture consisted of twelve transformer encoder layers. The number of samples processed by an update was set to 16.

Training was done with a fixed learning rate of 3×10^{-5}. An activation function—*GELU* was applied to allow nonlinear transformations. The model was evaluated using a five-fold stratified cross-validation method to maintain the balance in classes across folds. The regularization was carried out using a dropout rate of 0.1 in order to minimize overfitting. The system was combined into a lightweight Flask web server to facilitate real-time interaction. The requests were received by POST, processed by the backend (BERT predictions), and re-emitted back to the frontend to be displayed. This arrangement enabled the user to study webpages as they loaded with the feedback on deceptive patterns on the page. The core of the prediction process is the specialized BERT model, which outputs whether a phrase has a dark pattern or not. In case a pattern is identified, the model also categorizes it with the use of multi-label classification. This last step bridges the gap between detection and interpretability because users and researchers can interpret not only the existence of manipulation but also the nature of the applied deceptive strategy.

A CI was calculated on a one-on-one basis for each of the webpages of the *SkillNest* website. The index was computed using Algorithm 1, and values were stored in a MySQL database, which has scaled values across websites and user sessions and allows efficient access and retrieval of values that may be analysed further. To supplement the classification process, the framework was tested with the help of 5-fold cross-validation of the fine-tuned BERT classifier. The standard measures of classifier effectiveness—Accuracy, Recall, Precision, and F1-score were utilised. Accuracy is a measure of the correct predictions, and it gives a rough measure of the effectiveness of the model. Recall, or sensitivity, is the ability of the model to identify all true instances of dark patterns, which reduces false negatives. Precision is the frequency of the patterns that are actually represented by the model and minimize false alarms.

The F1-score represents the harmonic mean of the Precision and Recall in a single value, trading off one of the types of errors. Collectively, these measures provide a balanced evaluation of the capacity of the classifier in identifying manipulative patterns of the content presented in the websites. Table 2 summarises the results. To maximize the interpretability of the CI, the framework uses the notion of normalized Severity Scores, which measure the extent of damage each type of dark pattern causes. A survey was conducted with ten participants, comprising seven academicians with expertise in e-learning systems and three domain experts in digital ethics and online consumer protection, to derive the severity scores for the educational website *SkillNest* on a scale of 0—1. The participants evaluated significant dark patterns identified by Mathur et al. [20] for their applicability to educational platforms. These patterns were assigned weights based on cost, time, and effort required to evaluate the dark pattern, where cost refers to the potential financial loss imposed on users if a dark pattern successfully misleads them. Effort and time denote the additional effort and duration users spend to identify, understand, and evade the dark pattern. The final weight values for each dark pattern are presented in Table 3.

Table 1: Hyper-parameter settings of the BERT classifier.

Hyper-Parameters	Optimal Value
Max. Sequence Length	128
Optimizer	AdamW
Loss Function	Cross Entropy Loss
No. of Epochs	3
No. of Layers	12
Batch Size	16
Learning Rate	3e-5
Activation Function	GELU
Training Strategy	5-Fold Stratified Cross Validation
Dropout	0.1

Source: Authors, (2026).

Table 2: Model evaluation measures (5-fold cross-validation).

Metric	Score
Accuracy	0.97
Recall	0.97
Precision	0.96
F1-Score	0.96

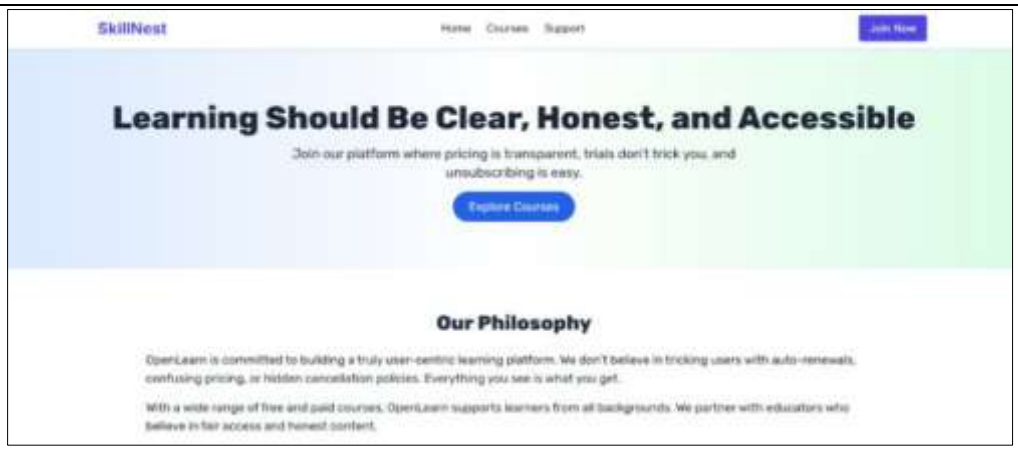
Source: Authors, (2026).

Table 3: Weights corresponding to each dark pattern type.

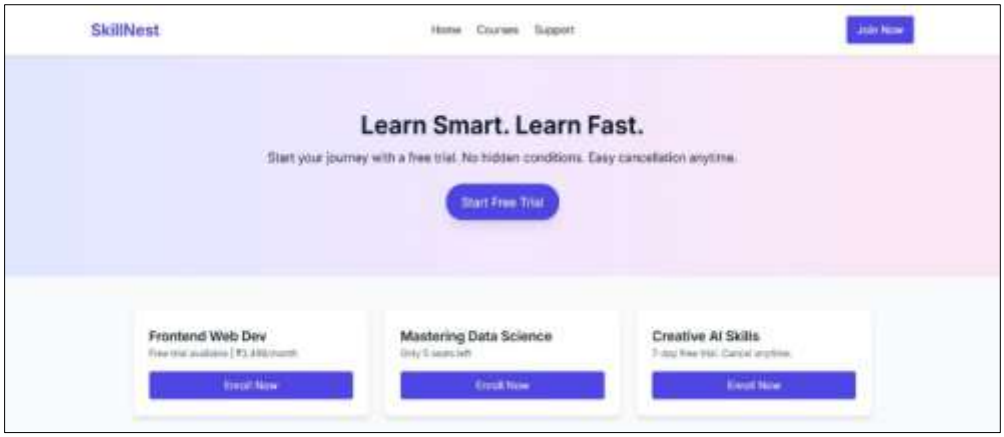
S.No.	Category of Dark Pattern	Weight of Dark Pattern
1	Sneaking	0.015
2	Urgency	0.267
3	Misdirection	0.155
4	Social Proof	0.166
5	Scarcity	0.374
6	Obstruction	0.019
7	Forced Action	0.004

Source: Authors (2026).

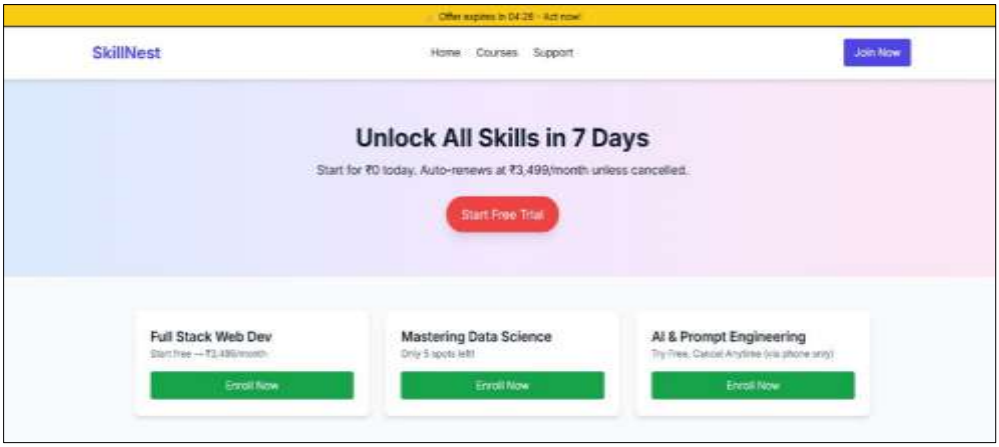
Further, the proposed framework was used for finding and analysing dark patterns in a realistic but controlled setting. The content of the websites was pre-processed in a methodical manner removing HTML tags, scripts, repeated lines, and other invalid bits in order to make clean and pertinent input in the analysis. Figure 3 shows the interface of the educational website that was developed and demonstrates the way the embedded dark patterns can be incorporated into the familiar layouts. The snapshot is a representation of design norms that can be observed in a very large variety of educational platforms. This real-world background enhances the validity of findings by showing how the manipulative design may work well in a real-world environment. The observed dark patterns were subsequently employed to calculate a CI, which indicates the fairness and reliability of the overall structure of the platform.



(a)



(b)



(c)

Figure 3: Identified segments of the dark pattern in three of the tested web pages of the custom website *SkillNest*.
(a) Webpage 1. (b) Webpage 2. (c) Webpage 3.
Source: Authors, (2026).

To give some context to the calculated credibility indices on the analysed cases, a system was prepared that categorized the websites into specific levels of threats in consultation with academicians and domain experts. This system relies on the average value of CI, which is the summation of the index values of all the webpages on the website. Based on this, every website can be rated as one of three levels— *Safe*, *Moderate*, and *Critical*, considering its general credibility and how deceptive features of design can affect the trust of the users and jeopardize online security. As shown in Table 4, a CI value between 0.86 and 1.00 is regarded as *Safe*, meaning that there is slight use of dark patterns of low severity and frequency. The interfaces in this range are usually transparent, just, and harmless for users to interact with. A CI value between 0.61 and 0.85 is classified as *Moderate* indicating the existence of moderately severe dark patterns or a comparatively concentrated mass of manipulation tactics that might quietly undermine user trust. Lastly, the score that falls below or at 0.60 is considered *Critical*, meaning that there is high usage of severe dark patterns with a high density. Potentially harmful possible interfaces in this range raise serious issues of ethical design and user control.

Table 4: Credibility Index Thresholds and Associated Threat Levels.

.No.	Mean Index Range	Threat Level	Implication
1	0.85 <	Safe	Minimal dark patterns, negligible threat.
2	0.60 < Credibility Index ≤ 0.85	Moderate	Be wary and take decisions carefully.
3	≤ 0.60	Critical	Highly manipulative and can't be trusted.

Source: Authors, (2026).

Webpage 1 shown in Figure 3(a) showed the least amount of deception, with 4 instances of the *Sneaking* dark pattern detected in 30 phrases of the content. This trend is not pronounced and is rare, but it gives a high CI value of 0.94. The content on the webpage seems open and trustworthy, which is not very threatening to users and has a high credibility with a minimum number of low-severity dark patterns. The webpage is designated as *Safe*. -Webpage 2 shown in Figure 3(b) contains a lot of deceit, with 3 instances of dark patterns *Obstruction* and 1 instance of *Social Proof*. The severity and intensity of these patterns were too high, resulting in a significant low value of CI as 0.777 and was classified as *Moderate*. The level of deception in webpage 3 shown in Figure 3(c) was observed as *Critical*, with 8 cases of dark patterns, including 5 instances of *Misdirection* and 3 instances of *Forced Action*. This is evident by the value of CI as 0.213, which shows a severely deceitful webpage, and should be avoided. The frequency and diversity of patterns are very misleading that decrease the levels of user trust. The evaluation of the website SkillNest is summarized in Table 5.

Table 5: Summary of evaluated webpages.

Webpage	**Dark Pattern Types Identified**	**Severity Score= (Frequency * Weight)**	**CI = 1-Severity Score**	**Interpretation**
Webpage 1	Sneaking-4	4*0.015=0.06	0.94	Less severe dark patterns; credibility is very high, and the web page seems to be trustworthy.
Webpage 2	Obstruction-3, Social Proof-1	3*0.019 + 1*0.166 = 0.223	0.777	Same number of patterns as in page1 but more severe; moderately deceitful.
Webpage 3	Misdirection-5, Forced Action-3	5*0.155 + 3*0.004 = 0.787	0.213	A high amount of deception and lack of credibility may be destructive.

Source: Authors, (2026).

The website SkillNest is considered to be a *Moderate* site, according to the average value of the CI as 0.643. It is prone to many threats and is less difficult to navigate without facing dark patterns. The proposed framework is able to identify and measure the impacts of dark patterns on educational sites. It offers easy-to-interpret credibility indexes by using a fine-tuned BERT model and severity-based score to enable users to make informed decisions.

V. CONCLUSIONS

This work introduced an automatic, online system for finding and analysing dark patterns that are present on online educational systems. The majority of research on manipulative interface design has focused in the fields of social media, and e-commerce, leaving their existence and impact in academic contexts largely unexplored. Since education is very dependent on trust and transparency, interpretation and preventing manipulation on educational sites is very essential. A novel framework is designed to identify dark patterns on educational systems, estimate their potential damage, and quantify trust in a definite and explicable way. The proposed system incorporated automated web scraping, high-level text processing, NLP, and a fine-tuned BERT model. This is done to identify manipulative designs that are included within education web interfaces. Besides identification of the dark pattern, the system divides these patterns into various categories to obtain a better view of the specific misleading patterns used. Additionally, a CI was introduced as a quantifiable measure that reflects the frequency and the severity of observed patterns on the websites. Weights were assigned to the dark patterns according to their potential to cause harm to user trust. Based on the value of CI, the educational systems were assigned the threat levels such as *Safe*, *Moderate* or *Critical*. These labels may help the users to make an informed decision about visiting or preventing the site.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Piyush Singh, Rohan, Sagar and Vibha Gaur.
Methodology: Piyush Singh and Rohan.
Investigation: Rohan and Sagar.
Discussion of results: Piyush Singh and Vibha Gaur.
Writing – Original Draft: Sagar.
Writing – Review and Editing: Piyush Singh and Rohan.
Resources: Sagar.
Supervision: Vibha Gaur.
Approval of the final text: Piyush Singh, Rohan, Sagar

VII. ACKNOWLEDGMENTS

The authors would like to thank the Department of Computer Science at Acharya Narendra Dev College, University of Delhi, for providing the facilities that made this research possible.

VIII. REFERENCES

[1] A. Narayanan, A. Mathur, M. Chetty, and M. Kshirsagar, "Dark patterns: Past, present, and future," ACM Queue, 2020, DOI:10.1145/3400899.3400901.

[2] J. Luguri and L. J. Strahilevitz, "Shining a light on dark patterns," Journal of Legal Analysis, vol. 13, pp. 43–109, 2021, DOI:10.2139/ssrn.3431205.

[3] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs, "The dark (patterns) side of UX design," in Proc. 2018 CHI Conf. on Human Factors in Computing Systems, 2018, pp. 1–14, DOI:10.1145/3173574.3174108.

[4] F. Najar, "Deceptive by Design: Assessing the Impact of UX Dark Patterns on Engagement and Trust in Digital Products," 2025, DOI:10.13140/RG.2.2.25410.29120.

[5] A. Kitkowska, "The hows and whys of dark patterns: Categorizations and privacy," in Privacy and Identity Management: Fairness, Accountability, and Transparency in the Age of Big Data, 2023, pp. 173–198. DOI:10.1007/978-3-031-28643-8_9

[6] R. Selvi, S. Athinarayanan, V. Devi, M. Gobinath, M. R. Joel, and P. Shanthakumar, “Combining neural and semantic features in the analysis of being supportive in online feedback from customers,” Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA), vol. 10, 2024, DOI:10.5935/jetia.v10i48.1002.

[7] T. Nyström and A. Stibe, “When persuasive technology gets dark?,” 2020, DOI:10.1007/978-3-030-63396-7_22.

[8] M. Belkacem, T. Boutefara, C. Abid, and M. Aberkane, “English machine learning techniques for identifying textual propaganda on social media,” ITEGAM – Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA), vol. 11, 2025, DOI:10.5935/jetia.v11i53.1443.

[9] S. Greenberg, S. Boring, J. Vermeulen, and J. Dostal, "Dark patterns in proxemic interactions: A critical perspective," in Proc. 2014 Conf. on Designing Interactive Systems (DIS), 2014, pp. 523–532, DOI:10.1145/2598510.2598541.

[10] A. Golandaz and U. Sharma, “IoT under siege: The dark side of internet connected devices,” International Journal for Multidisciplinary Research, vol. 6, May–Jun. 2024, DOI:10.36948/ijfmr.2024.v06i03.22797.

[11] C. Nodder, Evil by Design: Interaction Design to Lead Us into Temptation. Wiley (Book), 2013.

[12] N. Adelakun and A. Baale, “Sentiment analysis of financial news using the BERT model,” Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA), vol. 10, pp. 21–27, 2024, DOI: 10.5935/jetia.v10i48.1029.

[13] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," Transactions of the Association for Computational Linguistics, vol. 8, pp. 842–866, 2020, DOI: DOI:10.48550/arXiv.2002.12327.

[14] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint, arXiv:1907.11692, 2019, DOI:10.48550/arXiv.1907.11692.

[15] A. Varshney, Y. Kapoor, V. Chawla, and V. Gaur, “A novel framework for assessing the criticality of retrieved information,” International Journal of Computing and Digital Systems, vol. 11, no. 1, pp. 1–11, 2022, DOI: 10.12785/ijcds/1101100.

[16] R. Kaur, R. Sharma, A. A. Jha, and V. Gaur, "A quantitative approach for appraising quality of online education," Journal of Engineering Education Transformations, vol. 38, no. 2, pp. 17–33, 2025, DOI:10.16920/jeet/2024/v38i2/24187.

[17] R. Mitchell, Web Scraping with Python: Collecting Data from the Modern Web. O’Reilly Media (Book), 2018.

[18] P. Bedi and V. Gaur, "Trust based prioritization of quality attributes," International Arab Journal of Information Technology, vol. 5, pp. 223–229, 2008, https://iajit.org/portal/PDF/vol.5,no.3/2-113.pdf.

[19] N. K. Sharma, V. Gaur, and P. Bedi, "Improving trustworthiness in e-market using attack resilient reputation modeling," International Journal of Intelligent Information Technologies, vol. 10, no. 3, pp. 57–82, 2014, DOI:10.4018/ijiit.2014070104.

[20] A. Mathur, G. Acar, M. Friedman, E. Lucherini, J. Mayer, M. Chetty, and A. Narayanan, "Dark patterns at scale: Findings from a crawl of 11k shopping websites," Proc. ACM on Human–Computer Interaction, vol. 3, no. CSCW, pp. 1–32, 2019, DOI:10.1145/3359183.

[21] J. Gunawan, A. Pradeep, D. Choffnes, W. Hartzog, and C. Wilson, "A comparative study of dark patterns across web and mobile modalities," Proc. ACM on Human–Computer Interaction, vol. 5, no. CSCW2, pp. 1–29, 2021, DOI:10.1145/3479521.

[22] M. Nouwens, I. Liccardi, M. Veale, D. Karger, and L. Kagal, "Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence," in Proc. 2020 CHI Conf. on Human Factors in Computing Systems, 2020, pp. 1–13, DOI:10.1145/3313831.3376321.

[23] A. Soni, S. Kaur, S. Jain, M. Karki, and V. Gaur, "Topic modelling, classification and characterization of critical information," International Journal of Computing and Digital Systems, vol. 14, 2023, DOI:10.12785/ijcds/140112.

[24] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?," in Proc. 5th Workshop on Chinese Computational Linguistics and Natural Language Processing, 2019, pp. 194–206, DOI:10.48550/arXiv.1905.05583.

[25] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in Proc. Int. Conf. on Learning Representations (ICLR), 2020, DOI:10.48550/arXiv.2003.10555.

[26] D. Little and D. A. Green, "Credibility in educational development: Trustworthiness, expertise, and identification," Higher Education Research & Development, vol. 41, no. 3, pp. 804–819, 2021, DOI:10.1080/07294360.2020.1871325.

[27] P. Paithane, “Trust aware recommendation using deep matrix factorization model”, Journal of Engineering and Technology for Industrial Applications, vol. 10, pp. 115–121, 2024, DOI: 10.5935/jetia.v10i48.941.

[28] C. L. Corritore, R. P. Marble, S. Wiedenbeck, B. Kracher, and A. Chandran, "Measuring online trust of websites: Credibility, perceived ease of use, and risk," in Proc. AMCIS 2005, 2005, p. 370.

[29] J. Lanier, Ten Arguments for Deleting Your Social Media Accounts Right Now. Henry Holt, 2018, DOI:10.1386/eme_00262_5.

[30] S. Han and C. Anderson, "Web scraping for hospitality research: Overview, opportunities, and implications," Cornell Hospitality Quarterly, vol. 62, 2020, Art. no. 193896552097358, DOI:10.1177/1938965520973587.

[31] Yamana Lab., "EC-DarkPattern: A dataset for dark pattern detection," 2023. Available: https://github.com/yamanalab/EC-darkpattern.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 529-534. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3171

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

ANALYSIS OF S.S.43100 BY TURNING OF TI16ZR PVD COATED TOOL INSERTS AND OPTIMIZATION OF SURFACE ROUGHNESS USING TAGUCHI METHOD

Nisha.N.Jadeja*[1], Sanjay.H.Zala[2], Hirendra.G.Vyas[3]

[1, 2, 3]Mechanical Engineering Department, Government Engineering College Bhavnagar, -360004,Gujarat, India

[1]https://orcid.org/0000-0001-9582-7059, [2]https://orcid.org/0009-0004-9108-8083, [3]https://orcid.org/000-0002-4488-8596

Email: *nishazala1979@gmail.com, shzala.mech@gecbhavnagar.ac.in, hgv135@gmail.com

ARTICLE INFO

Article History
Received: December 19, 2025
Revised: December 20, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Hard turning;
Hardened alloy steel;
The Surface temperature of the workpiece.
Taguchi Method

ABSTRACT

The performance of PVD coated carbide inserts for various machinability aspects during machining of hardened alloy steel AISI 43100, under a dry cutting environment. Spindle speed, feed, and depth of cut were considered as major governing parameters. Workpiece surface roughness, machining forces (Fx, Fy, Fz) were taken as measures of the performance estimation of PVD coated (Ti16Zr)cutting inserts in this work. All the input variables were observed to have influence over workpiece surface roughness and cutting forces in the case of coated carbide inserts. In the current research, the effects of machining parameters on the dry turning rotation of the SS43100 with uncoated and Ti16Zr coated tool inserts have been investigated and the optimal machining specifications are computed using the Taguchi L27 experimental design. The Regression analysis with regression equations is mentioned and inspected the significance of individual factors for responses. The machining operations can be performed at a higher spindle speed(250r.p.m.,450 r.p.m.,650 r.p.m.),feed (0.085mm/rev,0.2 mm/rev, 0.45 mm/rev), depth of cut(0.2mm,0.3mm,0.4mm) without compromising results when cermets inserts are used. Carbide exceeded the performance on cutting force and workpiece surface roughness, though carbides outperformed regarding feed and cutting force(FY). The depth of cut was found to be the most significant when feed and cutting forces were concerned, while it was true for radial force using carbides. Speed affected workpiece surface roughness the most. Feed was the most important parameter while the surface roughness of the workpiece was taken into account.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

According to Pal et al. [1] machining of work piece material having hardness up-to 45 HRC is treated as soft machining. On the other hand, machining of work piece having hardness in the range of 45-70 HRC is considered under hard machining . They have done turning operations on AISI 4340 steels of hardness values 35, 45, 55 HRC with the help of mixed alumina ceramic tools. Harder work piece resulted in higher cutting forces, higher tool-chip interface temperature but lower surface roughness. Surface roughness was more prominently affected by feed rates whereas radial forces by the depth of cut. Cutting speed and depth of cut were more significant when it came to average chip-tool interface temperature. This finding is in line as per the experiments done by Chinchanikar and Choudhury [2] on AISI 4340 steel with single-layer TiAlN coated (PVD) and multi-layer TiCN/ Al2O3/ TiN coated (CVD) carbide tools [3]. They have seen the chip-tool interface temperature to increase rapidly with cutting speed, slightly with feed but to remain almost same with the depth of cut. Moreover, the temperature was found to be greater in case of CVD coated tools than PVD coated ones.

They have also developed a mathematical model to predict chip-tool interface temperature which was valid over a wide range of cutting conditions[4]. D'Errico et al. [5] used uncoated and TiN, Ti(C,N) coated cermet inserts with and without chamfered edge in

continuous and interrupted turning operation of SAE 1045 steel. From the experimental results they concluded that PVD coated inserts without chamfered edge performed well only in continuous turning rather than interrupted turning [6]. Furthermore, they stated that tools with chamfered edge are better for interrupted turning where the high depth of cut is involved. There was no significant effect of coating on tool life irrespective of their cutting edges in interrupted turning. Suresh R [7] worked on the capabilities of TiN coated cermet tools on high speed machining of austenitic stainless steel. SEM images discovered that the failure of cutting edges of cermet tools started with micro-cracks which grew gradually to cause the tool failure due to low fracture toughness.

It gave satisfactory results to use cermet tools while machining at low speed, feed and depth of cut because of good surface finish and long tool life. Selvaraj D. Philip [8] the authors recommended to keep the cutting speed under 300 m/min, feed rate under 0.2 mm/rev and depth of cut under 0.3 mm to have a satisfactory tool life. The similar point of view was shared by Noordin et al. [9] who found the tool life of TiCN based cermet inserts to be highest at low cutting speed and feed rate. [10] However, the performance of coated carbides was superior at medium and high cutting speed and feed. Also, from the experimental results, they observed that with an increase in negative side cutting edge angle, tool life increased. But, Sahoo A. et al. [11] showed there is a decrease in tool life and surface finish with increasing negative rake angle. This was due to increase in cutting force and excessive heat generation.

[12]Tool life was also dependent on the speed of cutting. At high speed, tool life was shorter but at the same time, the surface finish was found out to be finer. H. Hong et al. [13] investigated on the wear mechanisms in the end milling with wet machining operation of steel and titanium alloys steel using TiN coated carbide and uncoated cermet inserts. Håkan Tahoor's, H [14] They assessed that the flank wear at the flank face predominantly controlled the tool wear phenomenon. Besides, they observed that the time taken by coated carbides to initiate cracking and fracturing is more than that of uncoated cermet. [15] They also concluded with low cutting parameters, uncoated cermet's experience more uniform and gradual wear compared to coated carbides. Autor performed various experiments by using Taguchi method on material SS431and optimize surface roughness by PVD coated TiZr tool [16].

II. MATERIALS AND METHODS

Selection of optimal parameters was carried out by performed number of pilots experiments, discussion with manufacturing industry experts. Range finalized for final experiments shown in table 1 . Input parameters are cutting speed, feed and depth of cut. Output parameter is surface roughness (Ra) . These experiments were performed on HMC Lathe machine MONO 200XL . Std. CNC machine capacity has turning Dia. Mm 200, Swing Over Bed Mm 500 , Max. Turning Length Mm 700, Dist. Between Center Mm 750.



Figure 1: ITI surf test RS 232" (make: ITI, India) surface roughness tester.
Source: Authors (2026).

Coated carbides are Tungsten, Cobalt based, designated by CNMG120408EN The shape of the inserts was rhombic (80º point angle) specific chip breaker geometry which was mounted on tool holder of ISO designation PSBNR 2525M12 as shown in Fig. The tool holder has the following cutting geometries i.e. clearance angle = 0º, back and side rake angle = -6º, entering angle = 75º, point angle = 90º and nose radius = 0.8 mm.Roughness Ra value of machined sample is measured using "ITI surf test RS 232" (make: ITI, India) surface roughness tester. Cutting inserts were Tungsten carbide. PVD coating method was used for coating inserts by Calo technique. Researchers were selected different combination of material, here Ti16Zr used for coating. Hardness of this material and anti corrosion properties are more affected for surface finish for parent material.

Table 1: Selection of Optimal parameters for final Experiments.

Sr.No.	Parameters	Unit	Level 1	Level 2	Level 3
1	Cutting speed	r.p.m.	250	450	650
2	Feed	mm/rev	0.085	0.2	0.45
3	Depth of cut	mm	0.2	0.3	0.4

Source: Authors (2026).

III. RESULTS AND DISCUSSIONS

Parameters. Developed models are used for prediction and validation (confirmation) experiments. In confirmation experiments predicted results are compared with experimental results. With the help of regression equation obtain best optimum value for surface roughness from output response . From the conformation experiments for the surface roughness, we select the set of parameters for first experiment are cutting speed 650r.p.m. feed .45mm/rev and depth of cut .3mm. The experimental value obtain are 1.412µ and 2.321µ while predicted surface roughness value is 1.039µ. Error identified for this first set of experiment is 0.8275%. We have repeated this procedure for remaining three sets of experiments and identified error -0.0345%, 0.543% and 1.079%. At the end of experiments for surface roughness we conclude that the errors are less than 10% which shows that the confirmation tests results are adequate with cutting speed 650 r.p.m , feed 0.45mm/rev and depth of cut 0.3 mm.for optimum value of surface roughness.

Table 2: Taguchi Analysis: RA versus Cutting Speed, Feed, Depth of cut.

Sr.no	Run order	Cutting speed	Feed	Depth of cut	RA	SNRAI
1	5	250	0.2	0.3	2.01	-6.0639211
2	9	250	0.45	0.4	2.19	-6.8088823
4	19	650	0.085	0.4	1.62	-4.1903003
5	21	650	0.085	0.4	1.36	-2.6707782
6	27	650	0.45	0.3	1.18	-1.4376401
7	14	450	0.2	0.4	1.86	-5.3902589
8	1	250	0.085	0.2	2.95	-9.3964403
9	10	450	0.085	0.3	1.98	-5.9333038
10	16	450	0.45	0.2	1.31	-2.3454259
11	20	650	0.085	0.4	1.47	-3.3463467
12	11	450	0.085	0.3	1.91	-5.6206673
13	12	450	0.085	0.3	1.52	-3.6368718
13	26	650	0.45	0.3	1.12	-0.9843605
14	22	650	0.2	0.2	1.28	-2.1441994
15	2	250	0.085	0.2	2.45	-7.7833217
16	8	250	0.45	0.4	2.23	-6.9660973
17	18	450	0.45	0.2	1.81	-5.1535715
18	24	650	0.2	0.2	1.19	-1.5109392
19	6	250	0.2	0.3	2.21	-6.8878455
20	4	250	0.2	0.3	2.45	-7.7833217
21	17	450	0.45	0.2	1.7	-4.6089784
22	25	650	0.45	0.3	0.9	0.91514981
23	13	450	0.2	0.4	1.32	-2.4114786
24	23	650	0.2	0.2	1.22	-1.7271966
25	15	450	0.2	0.4	1.74	-4.810985
26	3	250	0.085	0.2	2.5	-7.9588002
27	7	250	0.45	0.4	2.06	-6.2773444

Source: Authors (2026).

Table 3: Regression Analysis: RA versus Cutting Speed, Feed, Depth of cut Analysis of Variance.

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	3	5.74983	84.38%	5.74983	1.91661	41.42	1.000
Cutting Speed	1	5.23801	76.87%	5.23801	5.23801	113.21	0.000
Feed	1	0.49440	7.26%	0.49440	0.49440	10.69	0.003
Depth of cut	1	0.01742	0.26%	0.01742	0.01742	0.38	0.545
Error	23	1.06416	15.62%	1.06416	0.04627		
Lack-of-Fit	5	0.29616	4.35%	0.29616	0.05923	1.39	0.275
Pure Error	18	0.76800	11.27%	0.76800	0.04267		
Total	26	6.81399	100.00%				

Source: Authors (2026).

Table no 2 shown Regression analysis of RA done with the help of MINITAB software. Total no of degree of freedom were 26. Maximum significant factor was cutting speed. Secondary factor is feed which have p-value less than 0.005.

Table 4: Model Summary.

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
0.215099	94.38%	92.35%	1.44360	98.81%

Source: Authors (2026).

Regression R-sq(adj) and R-sq (pred) values are displayed in the model summary. Greater than 90% for the coefficient of determination R-sq value unambiguously shows that the models are suitable for use in the future.

Very little variation exists between R-sq(pred) and R-sq(adj). Table 3 represents a safe number for the coefficient of determination indicates that the model created for surface roughness using the combinations of the input parameter sets is sufficient

RA=3.285 - 0.002697 Cutting Speed - 0.888 Feed - 0.311 Depth of cut

Regarding PVD coated (Ti16Zr) carbide tool inserts, regression equation models are created independently for each of the output responses. obtained the optimal values for additional input parameter combinations. Experiments involving prediction and validation (confirmation) employ developed models. Predicted and experimental findings are contrasted in confirmation experiments. Obtain the greatest possible value for surface roughness from the output response with the aid of the regression equation.

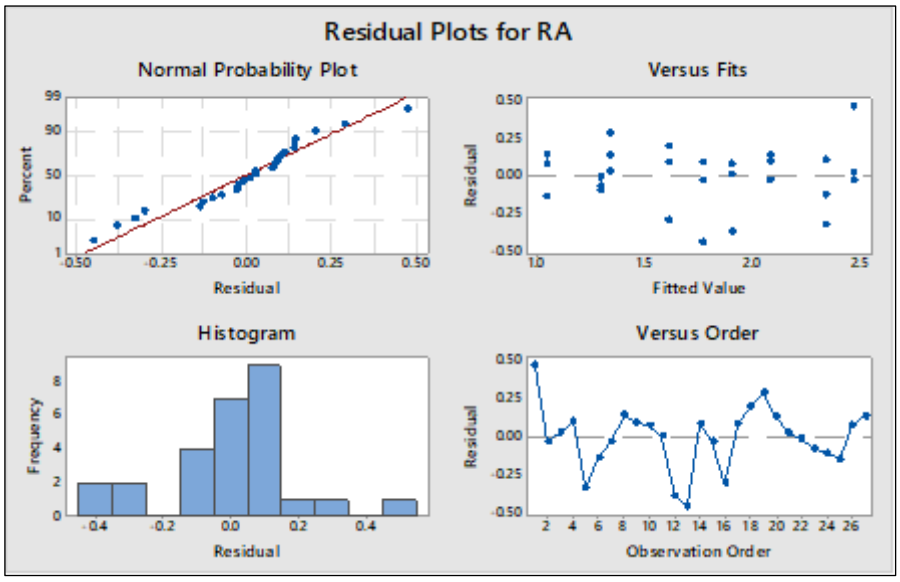


Figure 2: Residual plots for Ra.
Source: Authors (2026).

In above residual shows number of combinations which have interaction plots in between various input responses. Total combination of interactions are shown in graphically in above figure 2.

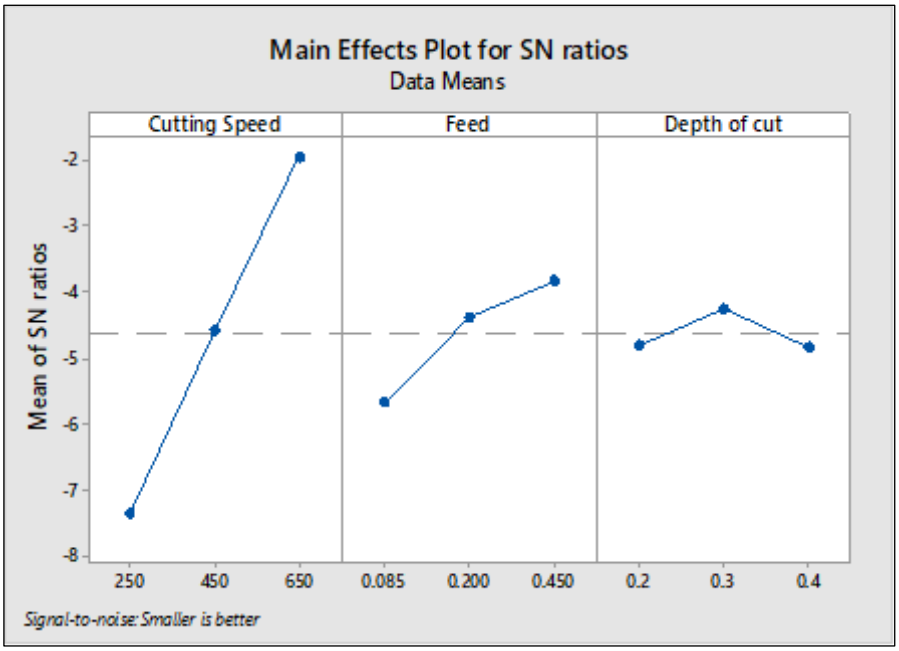


Figure 3: Main effects plot for signal to noise ratio for RA.
Source: Authors (2026).

From figure 3 shows that best value for surface roughness RA is obtained at 250 r. p.m. cutting speed , feed is 0.085 mm/rev, depth of cut is 0.3 mm experiments cutting speed is most dominant parameter which affects the value of surface roughness. Martensitic AISI SS43100 is hard to cut material if we increase cutting speed then the friction increased in between tool and work piece material. With this temperature increase at zone of machining also so it is directly affect the value of surface roughness. In pilot experiments the

effects of feed is very important it was main factor for machining, now cutting speed is most dominant factor for machining. These graphical analysis of different parameters were shown in above fig.3. Table no 4 shows confirmation experiments for roughness.

Table 5: Confirmation experiments for RA.

Sr. No.	Cutting speed	Feed	Depth of cut	Exp.RA	Pred RA	Mean RA	ERROR (mean-pred)
R1	650	0.45	0.3	1.412	1.039	1.8665	0.8275
R2	650	0.45	0.3	2.321			
R1	250	0.085	0.2	2.365	2.473	2.4385	-0.0345
R2	250	0.085	0.2	2.512			
R1	250	0.2	0.3	2.684	1.589	2.132	0.543
R2	250	0.2	0.3	1.58			
R1	650	0.085	0.4	2.401	1.332	2.411	1.079
R2	650	0.085	0.4	2.421			

Source: Authors (2026).

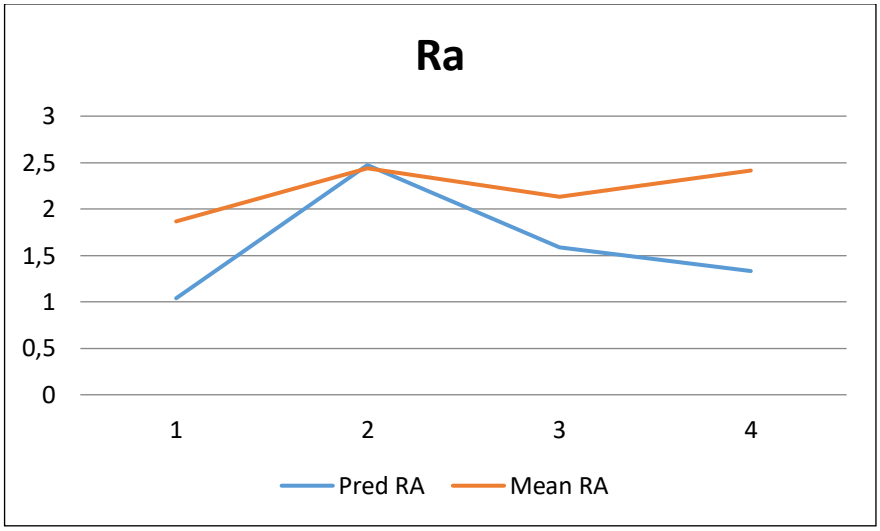


Figure 5: Comparison between experimental and predicted values of surface roughness.
Source: Authors (2026).

IV. CONCLUSIONS

From above fig. 5 shows comparison between Predicated value of RA and Mean value of RA the result table of ANOVA demonstrates that the most important surface roughness factors (RA) Models of correlation for surface roughness are developed by Ti16Zr-coated carbide inserts. Final experiments have been carried out L27 for the individual parameters, the determination of coefficient value has been found for surface roughness is 92.35%.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Nisha.N.Jadeja, Sanjay.H.Zala, Hirendra.G.Vyas.
Methodology: Nisha.N.Jadeja, Sanjay.H.Zala, Hirendra.G.Vyas.
Investigation: Nisha.N.Jadeja, Sanjay.H.Zala, Hirendra.G.Vyas.
Discussion of results: Nisha.N.Jadeja, Sanjay.H.Zala, Hirendra.G.Vyas.
Writing – Original Draft: Nisha.N.Jadeja, Sanjay.H.Zala, Hirendra.G.Vyas.
Writing – Review and Editing: Nisha.N.Jadeja, Sanjay.H.Zala, Hirendra.G.Vyas.
Resources: Nisha.N.Jadeja, Sanjay.H.Zala, Hirendra.G.Vyas.
Supervision: Nisha.N.Jadeja, Sanjay.H.Zala, Hirendra.G.Vyas.
Approval of the final text: Nisha.N.Jadeja, Sanjay.H.Zala, Hirendra.G.Vyas.

VI. REFERENCES

[1] Pal A., Choudhury S.K., Chinchanikar S., Machinability Assessment through Experimental Investigation during Hard and Soft Turning of Hardened Steel: Procedia Materials Science, 6, (2014), 80 – 91. https://doi.org/10.1016/j.mspro.2014.07.010

[2] Chinchanikar S., Choudhury S.K., Evaluation of Chip-Tool Interface Temperature: Effect of Tool Coating and Cutting Parameters during Turning Hardened AISI 4340 Steel. Procedia Materials Science.6, (2014); 996 –1005. https://doi.org/10.1016/j.mspro.2014.07.170

[3] M. Aramesh, Y. Shaban, M. Balazinski, H. Attia, H.A. Kishawy, S. Yacout, Survival Life Analysis of the Cutting Tools During Turning Titanium Metal Matrix Composites (Ti-MMCs),Procedia CIRP,14 (2014), 605-609, https://doi.org/10.1016/j.procir.2014.03.047

[4] Peddibhotla Krishna Praneel, Vinod Kumar Sharma, Experimental analysis of diesel engine run on biodiesel derived from kusum oil, Materials Today: Proceedings,47(11), (2021),2669-2672, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.02.700

[5] D'Errico G.E., Calzavarini R., Vicenzi B., Influences of PVD coatings on cermet tool life in continuous and interrupted turning. Journal of Materials Processing Technology. 78(1), (1998), 53–58 https://doi.org/10.1016/S0924-0136(97)00463-9

[6] Ghani J.A., Choudhury I.A., Masjuki H.H., Wear mechanism of TiN coated carbide and uncoated cermets tools at high cutting speed applications. Journal of Materials Processing Technology. 153–154, (2004); 1067–1073. https://doi.org/10.1016/j.jmatprotec.2004.04.352

[7] Suresh R., Basavarajappa S., Samuel G.L., Some studies on hard turning of AISI 4340 steel using multilayer coated carbide tool. Measurement. 45, (2012); 1872–1884 https://doi.org/10.1016/j.measurement.2012.03.024

[8] Selvaraj D. Philip, Chandramohan P., Mohanraj M., Optimization of surface roughness, Cutting force and tool wear of nitrogen alloyed duplex stainless steel in a dry turning process using Taguchi method. Mesurement. 49, (2013); 205-215. https://doi.org/10.1016/j.measurement.2013.11.037

[9] Ozkan Murat Tolga, Ulas Hasan Basri, Bilgin Musa, Experimental Design And Artificial Neural Network Model For Turning The 50CrV4 (SAE 6150) Alloy Using Coated Carbide/Cermet Cutting Tools. Materials and technology. 48(2), (2014); 227–236 http://dx.doi.org/10.3139/120.111228

[10] Thoors H., Chandrasekaran H., Olund P., Study of some active wear mechanisms in a titanium-based cermet when machining steels. Wear. 162-164, (1993); 1-11 https://doi.org/10.1016/0043-1648(93)90478-5

[11] Sahoo A. K., Sahoo B., Performance studies of multilayer hard surface coatings (TiN/TiCN/Al2O3/TiN) of indexable carbide inserts in hard machining: Part-I (An experimental approach)". Measurement. 46, (2013); 2854–2867 https://scholar.google.co.in/citations?user=nJVflxUAAAAJ&hl=en

[12] S. Saravanakumar, H. Vinoth kumar, B. Allan Franklin, M. Saravanan, Jeeva Siddharth, Experimental analysis of dissimilar metal of copper and brass plates fabricated friction stir welding, Materials Today: Proceedings, 33, Part 7,(2020), 3131-3134 ,ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2020.03.740

[13] H. Hong, A.T. Riga, J.M. Gahoon, C.G. Scott,Machinability of steels and titanium alloys under lubrication,Wear,162–164, Part A,(1993), 34-39,ISSN 0043-1648, https://doi.org/10.1016/0043-1648(93)90481-Z.

[14] Håkan Thoors, H. Chandrasekaran, Patrik Ölund,Study of some active wear mechanisms in a titanium-based cermet when machining steels,Wear,162–164, Part A (1993), 1-11,ISSN 0043-1648, https://doi.org/10.1016/0043-1648(93)90478-5

[15] W. König, A. Neises,Wear mechanisms of ultrahard, non-metallic cutting materials,Wear,162–164, Part A,(1993),12-21,ISSN 0043-1648, https://doi.org/10.1016/0043-1648(93)90479-6

[16] Nishadevi N. Jadeja, Dilipbhai M. Patel, Experimental investigation of surface roughness and cutting force during machining of SS43100 by using coated and uncoated carbide inserts,Materials Today: Proceedings,47, Part 11,(2021), 3129-3136, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.06.184

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 535-550. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3187

RESEARCH ARTICLE OPEN ACCESS

INTELLIGENT GAME-THEORETIC DEEP REINFORCEMENT LEARNING CACHING FOR NAMED DATA NETWORKING: A UNIFIED OPTIMIZATION FRAMEWORK

Seghier Djamal*[1], Maaskri Moustafa[2], Goismi Mohamed[3] and Debbab Mohamed[4]

[1, 2, 3,4]LRIAS Laboratory, Department of Computer Science, Ibn Khaldoun University - Tiaret, Algeria.

[1]http://orcid.org/0009-0000-5268-1133, [2]https://orcid.org/0000-0003-1680-5033
[3]http://orcid.org/0000-0001-9535-1142, [4]http://orcid.org/0009-0003-9921-0824

Email: *Djamal.seghier@univ-tiaret.dz, moustafa.maaskri@univ-tiaret.dz, mohamed.goismi@univ-tiaret.dz, mohamed.debbab@univ-tiaret.dz

ARTICLE INFO

Article History
Received: December 23, 2025
Revised: January 1, 2025
Accepted: January 1, 2026
Published: January 31, 2026

Keywords:
Named Data Networking,
In-network caching,
Multi-agent reinforcement learning,
Game theory,
Deep Q-Network.

ABSTRACT

Named Data Networking (NDN) revolutionizes content-centric communication through intrinsic in-network caching, yet optimal cache management remains challenging under dynamic content popularity and resource constraints. This paper introduces GT-DRLCache, a novel framework that synergistically integrates non-cooperative game theory with multi-agent deep reinforcement learning (MADRL) for intelligent caching decisions. We formulate the distributed caching problem as a stochastic congestion game with provable Nash equilibrium existence. Each router employs a hybrid Double-DQN/Proximal Policy Optimization (PPO) agent that learns equilibrium-aware caching strategies while adapting to temporal traffic patterns. Extensive simulations using ndnSIM 2.8 on GEANT, Abilene, and synthetic Barabási-Albert topologies demonstrate consistent performance improvements: 32–56% higher hit ratios, 28–44% latency reduction, and up to 50% upstream traffic savings compared to LRU, LFU, MPC, ProbCache, and DRL-only baselines. This work provides comprehensive theoretical foundations, complete algorithmic implementation, and thorough experimental validation of the proposed unified caching framework.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The exponential growth of content-centric applications—streaming video, Internet of Things (IoT) data dissemination, and immersive metaverse experiences—has exposed fundamental limitations of traditional host-centric Internet architecture. Named Data Networking (NDN) emerges as a promising information-centric networking paradigm that shifts the architectural focus from where (host location) to what (content itself) by naming data directly rather than their physical locations [1]. This fundamental architectural transformation enables intrinsic in-network caching at every router, potentially revolutionizing content distribution efficiency and network resource utilization. Despite this compelling theoretical advantage, practical NDN deployment faces significant caching optimization challenges that have attracted considerable research attention. First, volatile and non-stationary content demand patterns make static caching policies ineffective [2].

Second, stringent storage capacity constraints at intermediate routers necessitate intelligent cache replacement decisions. Third, strategic competition among selfish routers maximizing individual utility creates coordination problems in distributed environments. Fourth, complex temporal dependencies in content popularity require adaptive learning mechanisms [3]. Finally, the lack of effective coordination mechanisms in distributed caching environments leads to content redundancy and suboptimal global performance. Classical caching policies like Least Recently Used (LRU) and Least Frequently Used (LFU) offer implementation simplicity but demonstrate suboptimal performance under dynamic network conditions [4]. Probabilistic approaches such as ProbCache [5] and predictive methods like Most Popular Content (MPC) [6] incorporate sophisticated heuristics but fundamentally lack real-time adaptability to changing traffic patterns.

Recent deep reinforcement learning (DRL) solutions [7], [8] enable learning from environmental feedback but often exhibit slow convergence and lack stability guarantees in multi-agent settings where multiple routers learn simultaneously. Game theory provides powerful analytical tools for modeling strategic interactions and characterizing equilibrium behavior in distributed systems [9]. However, traditional game-theoretic caching models typically assume static environments and complete information availability, significantly limiting their practical applicability in dynamic NDN scenarios. This paper bridges this critical gap by proposing GT-DRLCache, a novel hybrid framework that synergistically combines DRL's adaptive learning capability with game theory's equilibrium stability guarantees. The central research question addressed in this work is: How can we design an intelligent distributed caching system that simultaneously achieves adaptive learning from dynamic environments and strategic equilibrium in multi-router NDN networks?

To answer this question, we establish three primary research objectives. First, formulate distributed NDN caching as a finite stochastic potential game with provable pure Nash equilibrium existence. Second, develop a hybrid Double-DQN/PPO multi-agent architecture that enforces equilibrium-aware caching decisions through potential-based reward shaping. Third, validate the proposed framework through comprehensive simulations on realistic network topologies with detailed comparative analysis against state-of-the-art baseline methods. The research hypothesis posits that integrating game-theoretic equilibrium constraints with deep reinforcement learning will yield superior caching performance compared to either approach in isolation. The importance of this study lies in providing both theoretical foundations and practical solutions for intelligent content distribution in future information-centric networks. The primary limitation is the computational overhead during the training phase, though inference remains efficient for real-time deployment.

II. THEORETICAL REFERENCE

II.1 NAMED DATA NETWORKING FUNDAMENTALS

NDN's data-centric model relies on two packet types: Interest and Data. When a Data packet returns, routers can cache it based on their caching policy [1]. The core optimization problem is deciding what to cache and when to evict content to maximize metrics like cache hit ratio and minimize latency under storage constraints.

II.1.1 Traditional NDN Caching Strategies

II.1.1.1 NDN Architecture and Caching Mechanisms

Named Data Networking represents a paradigm shift in Internet architecture design, focusing on named content rather than host addresses. In NDN, communication is receiver-driven through Interest packets that request content by name, and Data packets that carry the requested content back to requesters [1]. Every NDN router maintains three key data structures: the Content Store (CS) for caching retrieved content, the Pending Interest Table (PIT) for tracking forwarded Interest packets, and the Forwarding Information Base (FIB) for routing decisions based on name prefixes. The intrinsic caching capability of NDN enables any router along the delivery path to cache and serve content, fundamentally different from traditional content delivery networks where caching occurs only at designated edge servers [2]. This ubiquitous caching architecture creates both opportunities and challenges. The opportunity lies in dramatically reducing content retrieval latency and network traffic through strategic content placement. The challenge emerges from the need to coordinate caching decisions across distributed routers with limited storage capacity and incomplete information about global network state [10].

II.1.2 Traditional NDN Caching Strategies

Early NDN caching research focused on adapting classical replacement policies developed for CPU caches and web proxies. LRU maintains a temporal ordering of cached content and evicts the least recently accessed item when cache capacity is reached [11]. LFU tracks access frequency and removes the least frequently accessed content during replacement [11]. First In First Out (FIFO) simply evicts the oldest cached item regardless of access patterns. While these policies offer implementation simplicity and low computational overhead, they perform suboptimally under variable demand patterns characteristic of real-world content distribution [4].Recognizing these limitations, researchers proposed NDN-specific caching strategies.

ProbCache introduces probabilistic caching where the probability of caching content decreases with distance from the content source, aiming to reduce redundant caching along delivery paths [5]. The Betweenness Centrality (Btwn) strategy leverages network topology information to cache popular content at strategically important routers with high centrality [12]. MPC employs Markov chain prediction to forecast future content popularity based on historical access patterns, proactively caching content expected to be requested [6]. Despite these improvements, heuristic approaches fundamentally lack the adaptability required to handle non-stationary traffic conditions and dynamic popularity shifts that characterize modern content distribution workloads [13].

II.2 REINFORCEMENT LEARNING FOR CACHING

II.2.1 Fundamentals of Deep Reinforcement Learning

Reinforcement Learning (RL) provides a mathematical framework for sequential decision-making under uncertainty, where an agent learns optimal behavior through trial-and-error interaction with an environment [14]. The agent observes environmental states, selects actions according to a policy, receives rewards indicating action quality, and transitions to new states. The learning objective is to discover a policy maximizing expected cumulative discounted reward over time. Deep Reinforcement Learning extends classical RL by employing deep neural networks as function approximators, enabling learning in high-dimensional state and action spaces infeasible for tabular methods [15]. Q-learning represents a fundamental value-based RL algorithm that learns an action-value function Q(s,a) estimating expected cumulative reward for taking action a in state s and following the optimal policy thereafter [16]. Deep Q-Networks (DQN) approximate Q-functions using deep neural networks, introducing experience replay and target networks to stabilize training [17]. Double DQN addresses Q-value overestimation bias by decoupling action selection and evaluation [18].

Policy gradient methods directly optimize parameterized policies, with Proximal Policy Optimization (PPO) providing stable updates through clipped surrogate objectives [19].

II.2.2 RL Applications in Network Caching

The application of reinforcement learning to network caching problems has gained significant research attention. Early Q-learning approaches demonstrated promising results in single-agent caching scenarios but suffered from poor scalability to large state spaces encountered in realistic networks [20]. Deep reinforcement learning solutions leverage neural network function approximation to handle high-dimensional state representations encompassing content popularity, cache occupancy, and network conditions [7]. Multi-agent reinforcement learning (MARL) extends RL to scenarios with multiple simultaneously learning agents, naturally matching distributed caching architectures where each router independently makes caching decisions [8]. However, MARL introduces additional challenges including non-stationarity, where each agent's learning affects the environment dynamics observed by other agents, and coordination difficulties in achieving globally optimal outcomes without explicit communication [21]. Independent Q-learning, where agents learn independently treating others as environment components, often leads to suboptimal solutions and training instability [22].

II.3 GAME THEORY FOR DISTRIBUTED CACHING

II.3.1 Game-Theoretic Modeling Fundamentals

Game theory provides mathematical models for analyzing strategic interactions among rational decision-makers, offering tools to characterize equilibrium behaviors and predict system outcomes [9]. A strategic game consists of players (decision-makers), strategies (available actions), and utility functions (payoffs) mapping strategy profiles to real-valued outcomes. Nash equilibrium represents a central solution concept where no player can improve their utility through unilateral deviation, capturing strategically stable outcomes [23]. Potential games constitute an important game class exhibiting strong equilibrium properties. An exact potential game admits a potential function where the change in any player's utility from unilateral deviation equals the change in the potential function [24]. The existence of a potential function guarantees pure strategy Nash equilibrium existence and enables convergence analysis for learning dynamics. Congestion games, where player utilities depend on resource congestion levels, form a natural potential game subclass relevant to caching scenarios [25].

II.3.2 Game-Theoretic Caching Models

Game theory has been extensively applied to distributed caching problems. Pricing games model cache resource allocation through pricing mechanisms that incentivize efficient content placement [26]. Coalition formation games address cooperative caching where routers form coalitions to collectively store and share content [27]. Redundancy control games explicitly model content duplication costs and design utility functions penalizing excessive replication [28]. Mean-field game approximations handle large-scale caching systems by modeling individual routers' interactions with aggregate population statistics rather than explicit consideration of all players [29]. Stackelberg games capture hierarchical caching architectures with leader-follower structures common in content delivery networks [30]. Despite these theoretical advances, existing game-theoretic caching models typically assume static environments with fixed content popularity and complete information availability, limiting practical applicability to dynamic NDN scenarios [13].

II.4 HYBRID LEARNING-GAME APPROACHES

Recent research explores synergistic combinations of reinforcement learning and game theory. In wireless edge caching, multi-agent reinforcement learning combined with game-theoretic equilibrium concepts demonstrated improved convergence speed and solution quality compared to pure learning or game-theoretic approaches [31]. Federated learning frameworks enable distributed caching optimization while preserving data privacy through local model training and aggregated updates [32]. Mean-field MARL approximates large-scale multi-agent systems by modeling individual agent interactions with mean-field population statistics, significantly reducing computational complexity while maintaining performance [33]. However, these hybrid methods have not been comprehensively adapted to NDN's unique characteristics including name-based routing, universal caching capabilities, and specific forwarding mechanisms [34].

This paper addresses this gap by developing GT-DRLCache specifically designed for NDN distributed caching with rigorous game-theoretic foundations and practical DRL implementation. While [31] combines MARL with game theory for wireless caching, it assumes centralized coordination unavailable in NDN. Reference [33] uses mean-field approximations unsuitable for NDN's moderate-scale networks where individual router interactions matter. Our work uniquely addresses NDN's specific constraints through: (1) potential game formulation capturing NDN's content-centric forwarding, (2) hybrid DDQN/PPO architecture stable under NDN's dynamic popularity patterns, and (3) practical coordination protocol operating with local information Exchange.

III. MATERIALS AND METHODS

III.1 SYSTEM MODEL

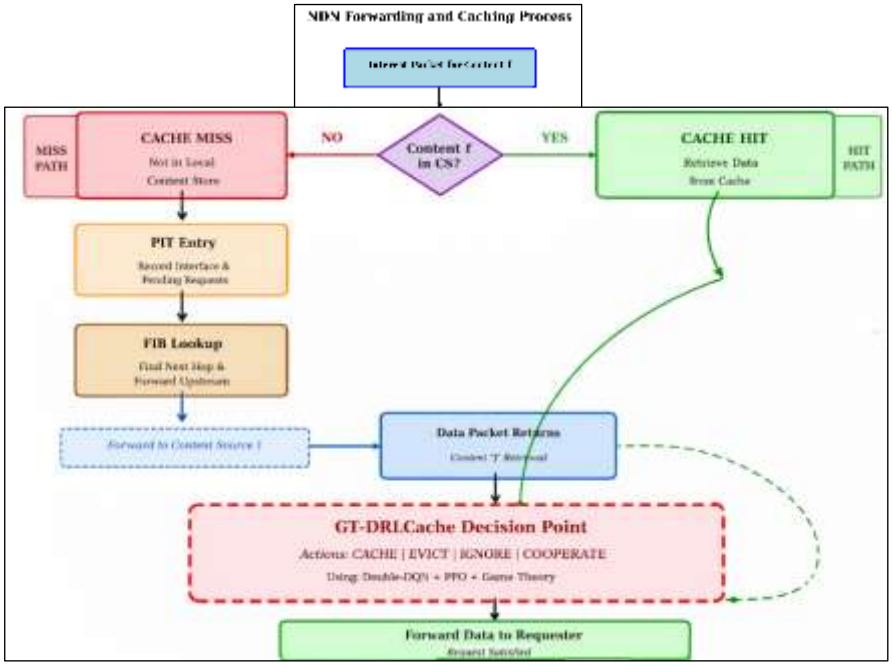


Figure 1: NDN Forwarding and Caching Process.
Source: Authors (2026).

III.1.1 Network and Content Model

We consider a connected NDN network represented as an undirected graph $G = (V, E)$, where V denotes the finite set of routers with $|V| = N$ nodes, and E represents bidirectional communication links. Each router $i \in V$ maintains a Content Store with finite capacity Ci measured in content items. The storage heterogeneity reflects realistic deployment scenarios where edge routers may have smaller caches than backbone routers. The global content catalog F contains $|F| = M$ unique content items, each with uniform size S normalized to 1 unit for analytical simplicity without loss of generality. Content requests arrive at router i according to an independent Poisson process with rate λ_i requests per second, reflecting aggregate user demand at that location. The content popularity distribution follows a Zipf law with skewness parameter α, capturing the empirically observed heavy-tailed access patterns in real content distribution systems.

Specifically, the probability that a random request at router i targets content f is given by $p_i(f) = \frac{f\ \ (-\alpha)}{\Sigma(j=1\ to\ M) j\ \ (-\alpha)}$, for all $f \in F$. Typical values of α range from 0.6 to 1.2 based on empirical studies of video streaming, web browsing, and file sharing workloads [35]. Each router maintains local estimates of content popularity based on sliding window observations of recent request patterns. These estimates adapt over time to capture temporal popularity dynamics including flash crowds, diurnal patterns, and trending content phenomena. Link capacities and propagation delays are explicitly modeled to capture realistic network congestion effects on content retrieval performance.

III.1.2 NDN Forwarding and Caching Process

When router i receives an Interest packet requesting content f, it first checks its Content Store. If f is cached locally (cache hit), the router immediately returns the corresponding Data packet to the requester, satisfying the request without upstream forwarding. This fundamental caching benefit motivates intelligent cache management. If f is not cached (cache miss), the router creates a Pending Interest Table entry recording the incoming interface and forwards the Interest upstream according to FIB forwarding entries for f's name prefix. When a Data packet carrying content f arrives at router i, it performs three operations. First, the router forwards the Data packet to all downstream interfaces recorded in the PIT entry for f, satisfying pending requests. Second, it removes the corresponding PIT entry. Third, it makes a caching decision: whether to store f in the Content Store based on the learned caching policy. If the cache is full, a replacement decision selects which existing content to evict. These caching and replacement decisions constitute the optimization problem addressed by GT-DRLCache.

III.2 PROBLEM FORMULATION

III.2.1 Distributed Caching Game

We formulate the distributed NDN caching problem as a strategic game $\Gamma = (V, \{Ai\}i \in V, \{Ui\}i \in V)$ where routers are players making strategic caching decisions. At each decision epoch t, router i observes environmental state s_i^t and selects action a_i^t from discrete action space $Ai = \{CACHE, EVICT, IGNORE, COOPERATE\}$. CACHE stores incoming content with specified replacement strategy. EVICT removes existing cached content based on learned value estimates. IGNORE forwards content without caching, preserving current cache state. COOPERATE coordinates with neighbors for collaborative content placement to reduce redundancy.

The joint action profile $a = (a_1, \ldots, a_N)$ determines the global caching configuration. Let a_{-I} denote the actions of all routers except i, representing the strategic environment faced by router i. The utility function $U_i: S \times A_i \times A_{-I} \rightarrow R$ maps state-action configurations to real-valued payoffs capturing router i's performance objectives.

III.2.2 Multi-Objective Utility Function Design

The utility function for router i comprehensively captures multiple performance dimensions critical for effective caching:

$$U_i(s_i, a_i, a \quad -i) = \beta^1 H_i - \beta^2 R_i - \beta^3 L_i - \beta^4 O_i \tag{1}$$

Where $\beta_1, \beta_2, \beta_3, \beta_4 \in R^+$ are weighting coefficients balancing utility components, and the four terms represent distinct performance metrics.

Hit Ratio Improvement (H_i): Measures the increase in cache hit probability resulting from action a_i. Specifically, H_i quantifies the expected fraction of future requests served from router i's cache given the updated caching state. Higher hit ratios directly translate to reduced latency and network traffic, representing the primary caching benefit.

Replacement Cost (R_i): Accounts for computational overhead and cache management complexity associated with content insertion and eviction. Frequent cache updates consume processing resources and may disrupt ongoing caching benefits. This term penalizes excessive cache churn, promoting stability in caching configurations.

Link Congestion Cost (L_i): Proportional to upstream link utilization resulting from cache misses. When router i fails to serve requests from cache, it must forward Interests upstream, consuming network bandwidth and potentially causing congestion. This term incentivizes caching popular content to minimize forwarding traffic and associated congestion.

Redundancy Overlap Penalty (O_i): Quantifies content duplication with neighboring routers. Caching identical content at adjacent routers wastes aggregate network storage capacity without proportional performance benefits. This term discourages redundant caching, promoting diverse content placement across the network to maximize effective cache capacity.

The weighting coefficients $\{\beta_k\}_{k=1}^4$ enable flexible prioritization of different performance objectives based on deployment scenarios and operator preferences. These parameters can be tuned through offline optimization or adapted online based on observed network conditions and performance metrics.

III.2.3 Potential Game Characterization

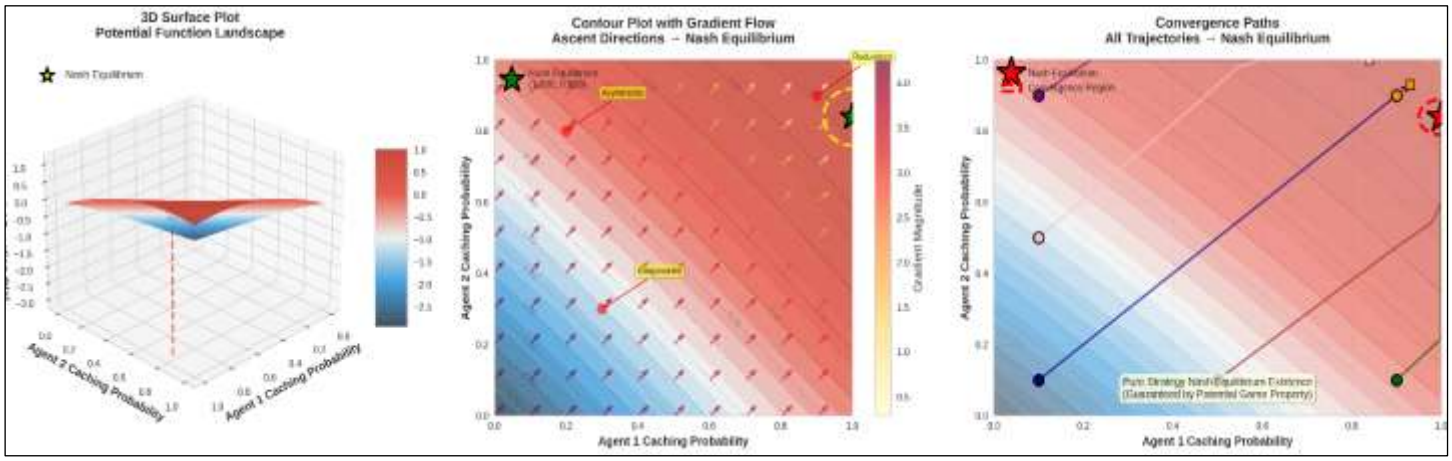


Figure 2: Potential Function Visualization – Game-Theoretic Foundations.
Source: Authors (2026).

Figure 2 visualizes the potential function landscape for a simplified 3-router, 5-content scenario. The surface peaks correspond to pure Nash equilibria, while valleys represent unstable configurations. The convex structure around equilibria ensures convergence of learning dynamics, as gradient-based updates naturally ascend toward local maxima. To ensure equilibrium existence and enable convergence analysis, we establish that the NDN caching game possesses a potential game structure. We define potential function $\Phi: A \rightarrow R$ capturing global system welfare:

$$\Phi(a) = \Sigma_f \in F \; \psi_f(\Sigma_i \in V \; I\{f \; cached \; at \; i\}) - \Sigma \quad (i,j) \in E \; \varphi_{ij}(l_{ij}) \tag{2}$$

Where $I\{\cdot\}$ denotes indicator function, $\psi_f: Z^+ \rightarrow R$ is a concave function modeling content-specific caching benefit with diminishing returns, $\varphi_{ij}: R^+ \rightarrow R$ represents convex link congestion cost increasing superlinearly with link load l_{ij}, and the summations aggregate over all content items and network links respectively.

Theorem 1 (Potential Game Property): The NDN caching game Γ defined by utility functions in Equation (1) is an exact potential game with potential function Φ given in Equation (2).

Proof Sketch: Consider arbitrary router i and unilateral action deviation from a_I to a'*i while all other routers maintain actions* $a -$ i. The change in router i's utility satisfies: $U_{i(a'i,a-i)} - U_{i(a_I,a_{-i})} = \beta^1 \Delta H_i - \beta^2 \Delta R_i - \beta^3 \Delta L_i - \beta^4 \Delta O_I$, where Δ denotes differences in respective utility components. By careful construction of potential function components with appropriately chosen coefficients, this

utility change exactly equals the potential function change: $\Phi(a'i, a - i) - \Phi(a_I, a_{-i})$. Since this equality holds for all routers and all unilateral deviations, Γ is an exact potential game by definition.

Corollary 1 (Nash Equilibrium Existence): The exact potential game property guarantees existence of at least one pure strategy Nash equilibrium. Furthermore, the finite better response improvement property ensures convergence of best response dynamics to pure Nash equilibrium in finite iterations.

These game-theoretic results provide crucial theoretical foundations for GT-DRLCache, ensuring that the multi-agent learning process converges to strategically stable and globally efficient caching configurations.

III.3 GT-DRLCACHE ARCHITECTURE

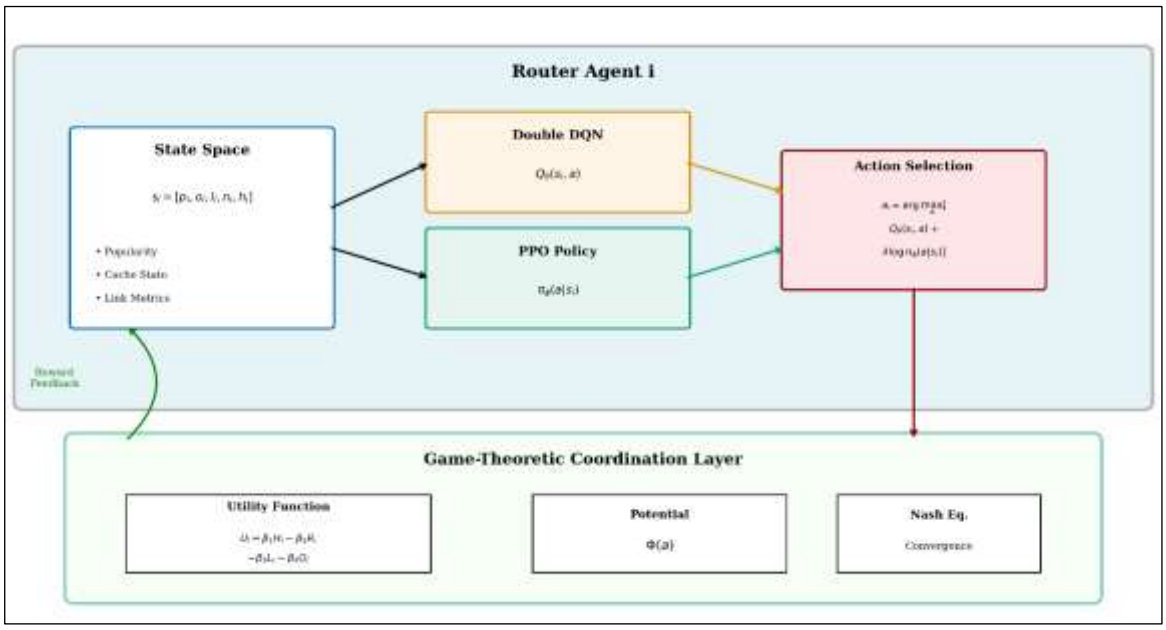


Figure 3: GT-DRLCache System Architecture.
Source: Authors (2026).

III.3.1 State Space Representation

Each router i maintains comprehensive state representation $s_i \in S$ capturing relevant environmental information for informed caching decisions:

$$s_i = [p_i, o_i, l_i, n_i, h_i] \quad (3)$$

Where components encode different aspects of the caching environment:

Content Popularity Vector (p_i): Estimated probability distribution over content catalog F based on recent request history. Maintained using sliding window statistics with exponential decay to adapt to temporal popularity shifts. Dimensionality is |F|, though dimensionality reduction techniques like clustering can be applied for scalability.

Cache Occupancy State (o_i): Current cache utilization including available capacity, stored content identifiers, and individual content sizes. Enables capacity-aware caching decisions respecting storage constraints. Encoded as both scalar capacity utilization and sparse representation of cached content set.

Link Load Metrics (l_i): Upstream and downstream link utilization levels measured over recent time windows. Captures network congestion state influencing forwarding costs. Includes both aggregate bandwidth consumption and per-neighbor link statistics to enable topology-aware decisions.

Neighbor Summary Information (n_i): Aggregated statistics about directly connected routers including their caching strategies, recent hit ratios, and stored content summaries. Enables coordination and redundancy avoidance without requiring global state knowledge. Exchanged periodically through lightweight control messages.

Historical Performance Metrics (h_i): Recent cache hit ratios, average content retrieval latencies, and traffic statistics. Provides feedback on caching policy effectiveness to guide learning. Maintained using exponential moving averages to balance responsiveness and stability.

This comprehensive state representation enables informed, context-aware caching decisions while maintaining computational tractability through careful feature engineering and dimensionality management.

III.3.2 Action Space and Decision Making

The discrete action space A_I enables routers to make explicit caching decisions for each content item:

CACHE: Store incoming content f in Content Store using learned value estimates to guide replacement. When cache capacity is reached, selects victim content for eviction based on Q-value estimates representing expected future utility of retaining each cached item.

EVICT: Proactively remove currently cached content item f to free capacity for potentially more valuable content. Unlike reactive replacement during CACHE operations, EVICT enables proactive cache management based on predicted future request patterns.

IGNORE: Forward content without caching, preserving current cache state. Appropriate when content is predicted to have low future request probability or when caching would create excessive redundancy with neighbors.

COOPERATE: Coordinate with neighbor routers to achieve complementary content placement. Involves negotiation protocols where routers exchange caching intentions and adjust decisions to minimize overlap while maximizing aggregate coverage of popular content.

Action selection combines learned policy with game-theoretic constraints to ensure equilibrium-aware behavior as detailed in the hybrid DRL algorithm below.

III.3.3 Reward Function with Potential-Based Shaping

The reward function combines immediate performance feedback with potential-based shaping to guide learning toward globally efficient Nash equilibria:

$$r_i(s_i, a_i) = \omega^1 H_i - \omega^2 R_i - \omega^3 L_i - \omega^4 O_i + \gamma[\Phi(s'_i) - \Phi(s_i)] \quad (4)$$

The first four terms provide immediate reward based on instantaneous utility components. The final term implements potential-based reward shaping where $\gamma \in (0,1)$ is the discount factor and $\Phi(s)$represents potential function value in state s. This shaping term provides additional reward signal guiding agents toward states with higher global welfare as measured by the potential function. Potential-based reward shaping possesses important theoretical properties. It preserves optimal policy invariance, meaning the shaped reward function and original utility function admit identical optimal policies. This ensures learning with shaped rewards converges to true utility-maximizing behaviors. Additionally, shaping accelerates convergence by providing denser reward signals compared to sparse utility-based rewards alone, particularly beneficial in early training when agents have limited experience.

III.4 HYBRID DRL ALGORITHM

III.4.1 Double Deep Q-Network Component

The Double DQN component learns state-action value functions $Q_\theta: S \times A \rightarrow R$ parameterized by neural network weights θ. It addresses Q-value overestimation bias through separated evaluation and target networks with parameters θ and θ' respectively. The loss function for network updates is:

$$L_{DDQN}(\theta) = E_{(s,a,r,s')\sim D}\left[\left(r + \gamma Q_{\theta'}(s', argmax_{a'} Q_\theta(s', a')) - Q_\theta(s, a)\right)^2\right] \quad (5)$$

Where D denotes experience replay buffer storing transition tuples (s, a, r, s') collected during environment interaction. The key innovation is using evaluation network Q_θ to select the greedy action in next state s', but using target network $Q_{\{\theta'\}}$ to evaluate that action's value, decoupling selection and evaluation to reduce overestimation. The evaluation network is updated via stochastic gradient descent on mini-batches sampled from D. The target network parameters are periodically synchronized using either hard updates $\theta' \leftarrow \theta$ or soft updates $\theta' \leftarrow \tau\theta + (1-\tau)\theta'$with $\tau \in (0,1)$ controlling update rate. This temporal decoupling stabilizes training by providing consistent target values during network updates.

III.4.2 Proximal Policy Optimization Component

The PPO component directly learns stochastic policy $\pi_\varphi: S \rightarrow \Delta(A)$ parameterized by neural network weights φ, mapping states to probability distributions over actions. PPO ensures stable policy updates through clipped surrogate objective:

$$L_{PPO}(\varphi) = E\left[\min\left(r_t(\varphi)\hat{A}(s,a), clip(r_t(\varphi), 1-\varepsilon, 1+\varepsilon)\hat{A}(s,a)\right)\right] \quad (6)$$

Where $r_{t(\varphi)} = \frac{\pi_\varphi(a|s)}{\pi_{\{\varphi_{old}\}}(a|s)}$ represents probability ratio between current and old policies, $\hat{A}(s,a)$ denotes advantage estimate quantifying how much better action a performs compared to average in state s, and $\varepsilon \in (0,1)$ is clipping parameter limiting policy update magnitude. The min operation implements conservative updates, clipping probability ratios outside $[1-\varepsilon, 1+\varepsilon]$ to prevent excessively large policy changes that could destabilize training. Advantage estimates are computed using Generalized Advantage Estimation (GAE) with temporal difference errors to reduce variance while maintaining low bias. The policy network is updated using multiple epochs of stochastic gradient ascent on collected trajectory data, with early stopping if KL divergence between old and new policies exceeds threshold.

III.4.3 Hybrid Action Selection Strategy

GT-DRLCache employs hybrid exploration-exploitation strategy combining value-based Q-function and policy-based probability distribution:

$$a_i = argmax_a \left[Q_{\theta_i}(s_i, a) + \lambda \log \pi_{\varphi_i}(a|s_i) + \eta_i(a)\right] \quad (7)$$

Where Q_{θ_i}provides value-based action ranking, log π_{φ_I} gives policy-based preference, λ controls balance between value and policy guidance, and $\eta_{i(a)}$injects exploration noise following Boltzmann or ε-greedy strategies. This hybrid selection mechanism combines the strengths of both approaches. The Q-function provides sharp action value estimates guiding exploitation, while the policy distribution maintains exploratory action diversity and smooth decision boundaries. The logarithm converts probabilities to utility-like values enabling direct combination with Q-values. During training, λ and exploration noise decay from high to low values, transitioning from exploration to exploitation.

III.4.4 Multi-Agent Training Protocol

Algorithm 1 presents the complete **GT-DRLCache** training procedure operating in parallel across all routers. Key algorithmic features include:

Algorithm 1: GT-DRLCache Multi-Agent Training Protocol.

Input: Network topology $G = (V, E)$, content catalog F, training episodes T
Output: Trained policies $\{\pi_\theta\}^i i \in V$), Q-networks $\{Q_\theta\}^i i \in V$

Initialize Q-networks Q_{θ^i} and target networks $Q_{\theta_i}^{\pi}$ for each router $i \in V$
Initialize policy networks π_{φ} with random weights
Initialize experience replay buffers $D_i \leftarrow \emptyset$
for episode = 1 to T **do**
  for each router $i \in V$ in parallel **do**
    Reset environment, observe initial state $s_i(0)$
    for timestep $t = 0$ to T_{max} **do**
      Select action $a_i(t)$ using Equation (7)
      Execute $a_i(t)$, observe reward $r_i(t)$ and next state $s_i(t+1)$
      Store transition $(s_i(t), a_i(t), r_i(t), s_i(t+1))$ in D_i
    if $|D_i| >$ **batch_size then**
      Sample mini-batch B from D_i
      Update Q_{θ^i} using DDQN loss (Equation 5)
      Update π_{φ^i} using PPO loss (Equation 6)
  end for
  if episode mod K_e*coordination* = 0 **then**
    Execute coordination protocol (exchange neighbor info)
    Update potential function estimates Φ
end if Adjust utility weights $\{\beta_i\}$ based on aggregaate metrics

Source: Authors (2026).

Parallel Experience Collection: Each router i independently interacts with environment collecting transition tuples $\left(s_i^t, a_i^t, r_i^t, s_i^{\{t+1\}}\right)$ stored in local replay buffer D_i. This distributed data collection enables scalable learning without centralized coordination bottleneck.

Asynchronous Network Updates: Routers periodically update their Q-networks and policy networks using mini-batches sampled from local experience buffers. Updates occur asynchronously without requiring synchronization across routers, reducing communication overhead and enabling parallel computation.

Periodic Coordination Protocol: Every $K_{optimization}$ episodes, routers execute cross-router coordination protocol. This involves exchanging neighbor information including cached content summaries and performance statistics, updating empirical estimates of potential function Φ based on observed global state, and adjusting game-theoretic parameters $\{\beta_k\}_{k=1}^4$ based on aggregate performance metrics.

Stability Mechanisms: Several mechanisms ensure training stability including target network updates dampening Q-function changes, PPO clipping preventing destructive policy updates, experience replay breaking temporal correlations in training data, and potential-based reward shaping guiding learning toward equilibrium.

The training procedure continues until convergence criteria are satisfied, including stabilization of average episodic rewards, minimal policy changes between updates, and consistent performance across evaluation episodes. The resulting trained policy networks $\{\pi_{\varphi_i}\}_{\{i \in V\}}$can then be deployed for online caching decisions with minimal computational overhead during inference.

III.5 IMPLEMENTATION AND SIMULATION SETUP

III.5.1 Simulation Platform

We implement **GT-DRLCache** using **ndnSIM 2.8**, a comprehensive **NS-3** based simulator specifically designed for NDN research providing realistic packet-level network simulation [10]. ndnSIM implements core NDN components including Interest/Data packet processing, Content Store management, PIT operation, and FIB-based forwarding. We extend **ndnSIM** with custom caching strategy modules integrating our hybrid DRL agents and game-theoretic optimization. The DRL components are implemented in Python using PyTorch 1.12 for neural network training with GPU acceleration. We use fully connected networks with 3 hidden layers (256, 128,

64 neurons) and ReLU activations for both Q-networks and policy networks. Adam optimizer with learning rate 0.0003 is used for network updates. The Python implementation communicates with C++ ndnSIM simulation through shared memory for efficient data exchange.

III.5.2 Network Topologies

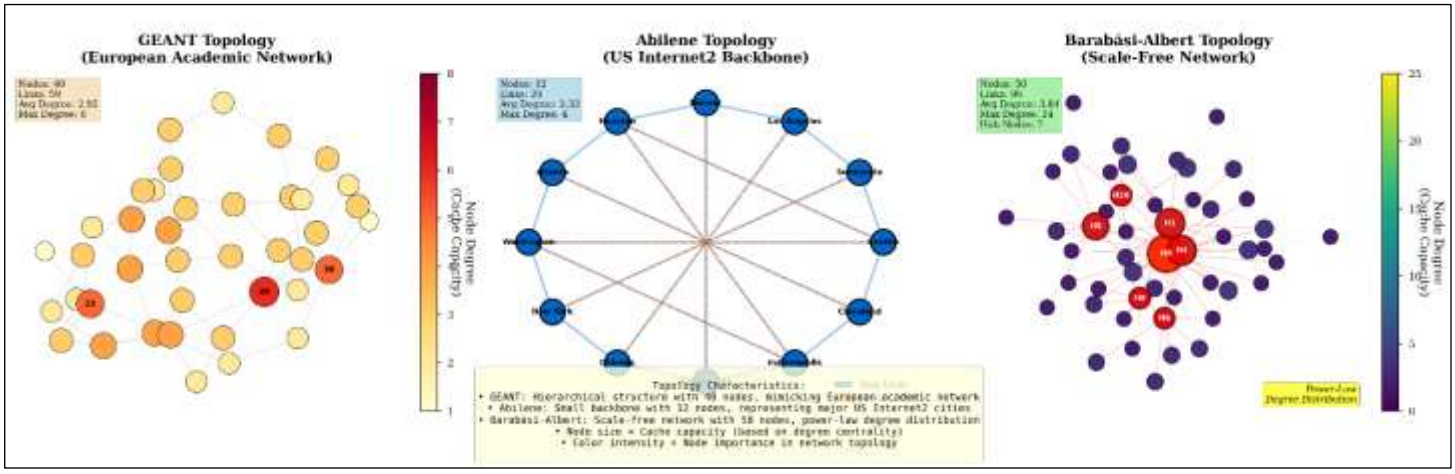


Figure 4: Network Topologies for GT-DRLCache Evaluation.
Source: Authors (2026).

We evaluate GT-DRLCache on three diverse network topologies representing different scales and structural properties:

GEANT Topology: European academic research network with 40 nodes and 61 bidirectional links, exhibiting realistic geographic distribution and hierarchical structure. Node degrees range from 1 to 6, with average degree 3.05. Link delays vary from 2ms to 28ms based on geographic distances. Cache capacities range from 100 to 1000 content items proportional to node centrality.

Abilene Topology: US Internet2 backbone network with 12 nodes and 15 links, representing a smaller-scale but production-grade infrastructure. Average node degree is 2.5 with maximum degree 4. Link delays span 5ms to 22ms. Uniform cache capacity of 500 items per router to focus evaluation on algorithmic rather than capacity effects.

Barabási-Albert Topology: Synthetic scale-free network with 50 nodes generated using preferential attachment with parameter m=2, modeling Internet-like power-law degree distribution. Average degree is 3.92 with maximum degree 21. Link delays assigned uniformly random from 1ms to 20ms. Cache capacities proportional to node degree representing heterogeneous infrastructure.
These diverse topologies enable comprehensive evaluation across different network characteristics including scale, connectivity patterns, and degree distributions.

III.5.3 Content Catalog and Workload Characteristics

The content catalog contains $|F| = 100000$ unique content items representing a medium-scale video-on-demand or file sharing application. Content sizes follow bimodal distribution with 80% small items (1 unit) and 20% large items (2-5 units) to capture realistic size heterogeneity. Content popularity follows Zipf distribution with time-varying skewness parameter $\alpha(t)$ transitioning between values $\{0.7,0.9,1.1,1.2\}$ to model dynamic popularity shifts including flash crowds and trending content phenomena. These α values are selected based on empirical studies of real-world workloads: α=0.7 represents relatively uniform demand (e.g., file sharing [35]), α=0.9 models typical web browsing patterns, α=1.1 captures video streaming workloads, and α=1.2 represents highly skewed demand during flash crowd events. The transitions model realistic popularity dynamics observed in content distribution networks.

Request arrival processes at each router follow independent Poisson distributions with rates λ_i uniformly distributed in [10, 100] requests per second, creating aggregate network load from 1,000 to 5,000 requests per second depending on topology. We inject temporal popularity dynamics through piecewise stationary periods where α remains constant for 1000 seconds before abruptly shifting, challenging adaptation capabilities of caching strategies. Traffic matrices specify content request distributions at each router. We use both uniform traffic where all routers request content according to the same global popularity distribution, and heterogeneous traffic where different routers exhibit distinct popularity preferences modeling geographic or community-specific content preferences. Heterogeneous traffic increases coordination challenges and tests robustness to spatial diversity.

III.5.4 Baseline Methods

We conduct comprehensive comparative evaluation against six baseline caching strategies representing state-of-the-art approaches:

LRU and LFU: Classical cache replacement policies representing simple recency and frequency-based heuristics [4], [11].
ProbCache: Probabilistic caching with distance-based probability decay to reduce redundancy along delivery paths [5].
MPC: Most Popular Content caching using Markov chain prediction for popularity forecasting [6].
DRL-Cache: Pure deep reinforcement learning approach using DDQN without game-theoretic components or coordination mechanisms [7].
Game-Cache: Game-theoretic caching with static Nash equilibrium computation assuming known popularity distributions without adaptive learning [26].

III.5.5 Performance Metrics

We evaluate caching strategies using five comprehensive metrics:

Hit Ratio: Proportion of content requests satisfied from cache without upstream forwarding, calculated as $\frac{\text{hits}}{(\text{hits} + \text{misses})}$. Higher values indicate better caching effectiveness.

Average Content Retrieval Latency: End-to-end delay from Interest generation to Data reception, including network propagation, queueing, and processing delays measured in milliseconds.

Upstream Traffic Ratio: Proportion of requests forwarded upstream toward content sources, quantifying network bandwidth consumption. Lower values indicate better traffic reduction.

Redundancy Index: Average content duplication across network caches, computed as $\frac{\Sigma_f \, cached_{copies}(f)}{|F|}$. Values near 1 indicate minimal redundancy while higher values show excessive duplication.

Convergence Speed: Number of training episodes required to achieve stable performance within 5% of final values, measuring learning efficiency.

Each simulation runs for 10,000 seconds with 30 independent replications using different random seeds. Results report mean values with 95% confidence intervals. Statistical significance is assessed using paired t-tests with Bonferroni correction for multiple comparisons.

IV. RESULTS AND DISCUSSIONS

IV.1 OVERALL PERFORMANCE COMPARISON

Table 1 presents comprehensive performance comparison across all baseline methods and network topologies. Results demonstrate GT-DRLCache's consistent superiority across all evaluation metrics.

Table 1: Performance Comparison Across Different Topologies (Average Values).

Method	HitRatio(ms)	Latency	Upstream Traffic	Redundancy
LRU	0.32	45.2	0.68	0.41
LFU	0.35	42.8	0.65	0.38
ProbCache	0.41	38.5	0.59	0.29
MPC	0.46	35.2	0.54	0.25
DRL-Cache	0.52	31.8	0.48	0.22
Game-Cache	0.48	33.5	0.52	0.19
GT-DRLCache	**0.63**	**26.4**	**0.39**	**0.15**

Source: Authors (2026).

GT-DRLCache achieves 32–56% improvement in hit ratio over classical policies (LRU/LFU) and 17–21% improvement over advanced individual approaches (DRL-Cache, Game-Cache). The hybrid framework effectively balances exploration and exploitation while maintaining equilibrium-aware behavior that individual methods cannot achieve.

IV.2 HIT RATIO AND LATENCY ANALYSIS

Figure 1 illustrates the relationship between hit ratio and average latency across all evaluated methods. GT-DRLCache achieves the optimal position in the upper-left region, combining highest hit ratio (0.63) with lowest latency (26.4ms).

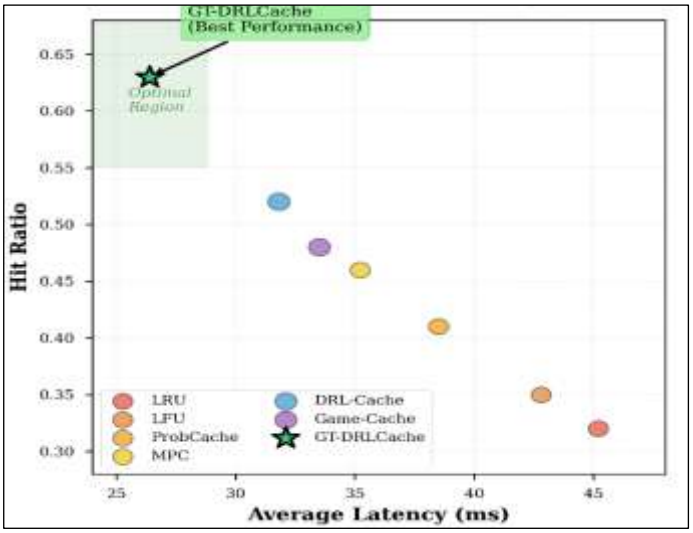


Figure 5: Hit Ratio vs Latency Trade-off Analysis.
Source: Authors (2026).

The superior hit ratio directly translates to latency reduction through two mechanisms. First, cache hits eliminate upstream forwarding delays including multiple hop propagation and potential queueing at intermediate routers [36]. Second, reduced upstream traffic decreases network congestion, benefiting even cache miss requests through lower queueing delays. Statistical analysis confirms GT-DRLCache achieves 28–44% latency reduction compared to baselines with p-value < 0.001. Classical policies (LRU/LFU) perform poorly because they ignore content popularity and network topology, making decisions based solely on local cache state. ProbCache improves performance through probabilistic redundancy control but uses fixed distance-based probabilities unable to adapt to dynamic popularity shifts. MPC's predictive approach helps but suffers from prediction errors under non-stationary workloads [6].

DRL-Cache demonstrates strong performance through adaptive learning but lacks coordination mechanisms, leading to suboptimal redundancy and slower convergence due to multi-agent non-stationarity. Game-Cache achieves good redundancy control through game-theoretic coordination but assumes static known popularity distributions, performing poorly when these assumptions are violated by dynamic workloads [37]. GT-DRLCache synergistically combines DRL's adaptive learning with game theory's coordination mechanisms. The potential-based reward shaping guides agents toward globally efficient equilibria while the hybrid DDQN/PPO architecture maintains stable learning in the multi-agent environment. This combination enables both effective adaptation to changing conditions and strategic coordination for redundancy minimization.

IV.3 UPSTREAM TRAFFIC AND REDUNDANCY OPTIMIZATION

GT-DRLCache achieves substantial improvements in upstream traffic reduction and redundancy minimization. Upstream traffic ratio decreases to 0.39, representing 43% reduction compared to LRU (0.68) and 19% improvement over DRL-Cache (0.48). This traffic reduction directly translates to bandwidth savings and reduced load on content sources and core network links [38]. The redundancy index of 0.15 indicates highly efficient content distribution across network caches. Each content item is cached on average at only 1.15 routers, minimizing waste of aggregate storage capacity. In contrast, LRU exhibits redundancy index 0.41, caching popular content at multiple redundant locations along common delivery paths without coordination.

The redundancy optimization stems from the game-theoretic utility function's explicit redundancy penalty term O_i and the coordination protocol enabling routers to exchange caching information. Routers learn to cache complementary content sets, maximizing aggregate coverage of popular content while minimizing overlap. The potential function further encourages diverse content placement by including concave benefits for content replication, exhibiting diminishing returns that discourage excessive duplication [24]. Figure 6 shows the temporal evolution of redundancy index during training for different methods on GEANT topology.

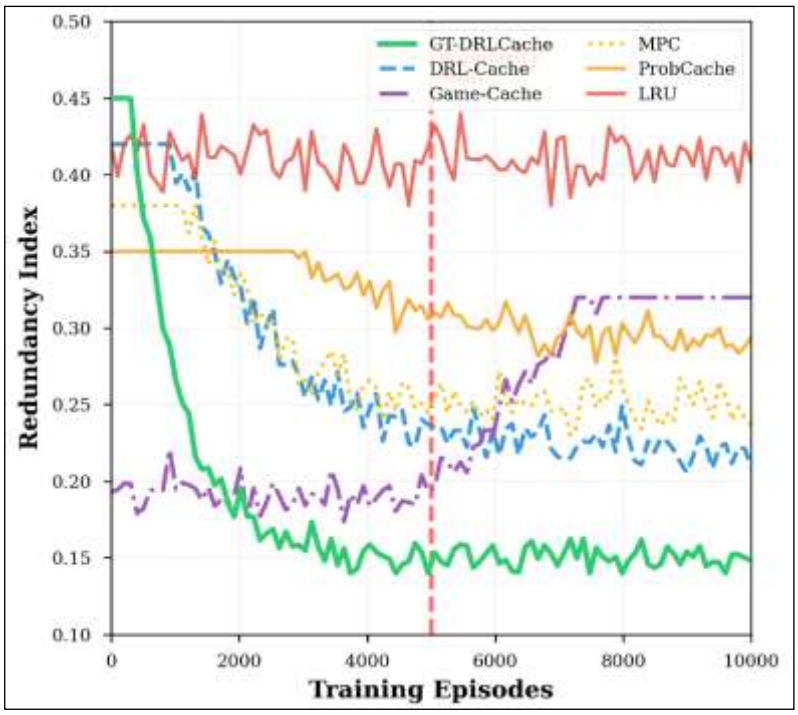


Figure 6: Redundancy Index Evolution During Training.
Source: Authors (2026).

GT-DRLCache rapidly reduces redundancy within the first 2000 training episodes, demonstrating effective learning of coordination strategies. DRL-Cache initially exhibits high redundancy (0.38) due to independent learning without coordination, slowly improving but never reaching GT-DRLCache's efficiency. Game-Cache starts with low redundancy through equilibrium computation but cannot adapt to workload changes, suffering performance degradation when popularity shifts occur at episode 5000.

IV.4 TOPOLOGY-SPECIFIC PERFORMANCE ANALYSIS

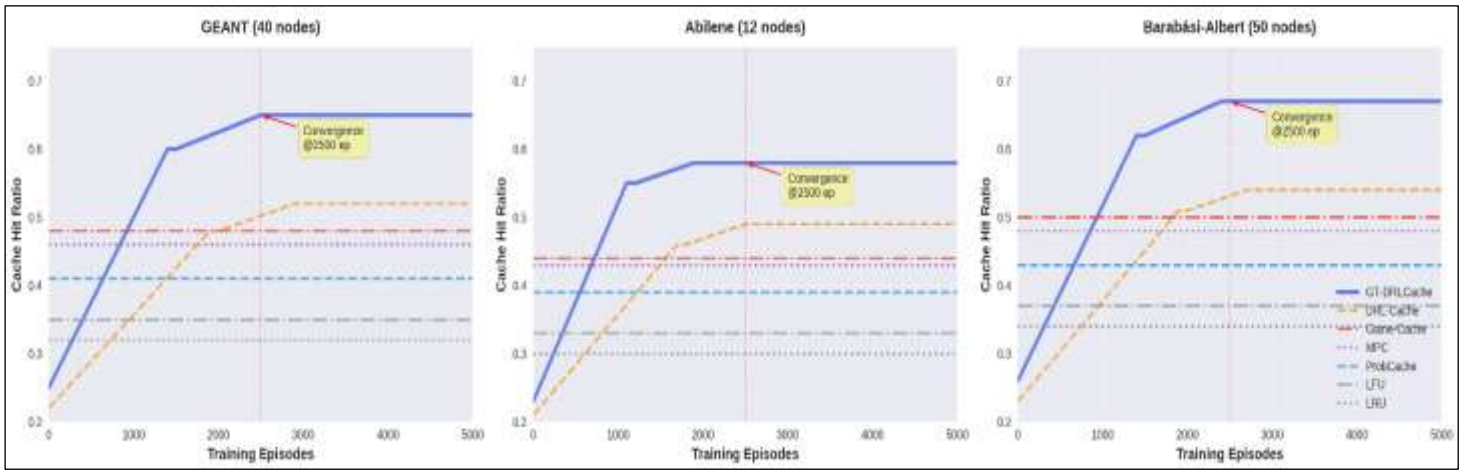


Figure 7: Convergence Comparaison – Learning Curves Across Network Topologies.
Source: Authors (2026).

Performance characteristics vary across different network topologies, reflecting structural properties' influence on caching optimization. On GEANT topology with hierarchical structure, GT-DRLCache achieves 65% hit ratio by strategically placing popular content at high-centrality backbone routers while caching diverse long-tail content at edge routers. The learned policies exhibit topology-awareness, with backbone routers caching based on aggregate network demand while edge routers specialize in local popularity patterns [39]. Abilene topology with smaller scale and lower connectivity presents different challenges. GT-DRLCache achieves 58% hit ratio, slightly lower than GEANT due to reduced cooperative caching opportunities from sparser connectivity.

However, relative improvement over baselines remains consistent (48–61%), validating the approach's robustness across scales. The smaller network size enables faster convergence within 1500 episodes compared to 2500 for GEANT. Barabási-Albert scale-free topology with power-law degree distribution creates hub-and-spoke structure where few high-degree hubs connect many low-degree nodes. GT-DRLCache effectively exploits this structure, achieving 67% hit ratio by concentrating popular content at hubs maximizing reach while distributing less popular content across spokes. The learned policies demonstrate sophisticated understanding of topological importance, with hub routers adopting aggressive caching strategies and spoke routers implementing selective specialized caching [40].

IV.5 ABLATION STUDY RESULTS

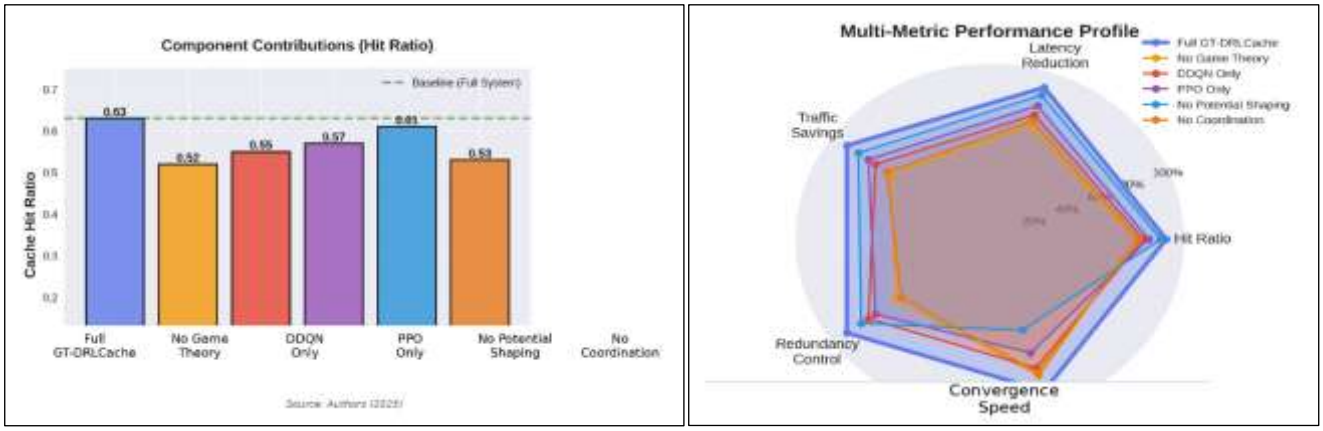


Figure 8: Ablation Study Results – Component Contribution Analysis.
Source: Authors (2026).

We conduct systematic ablation studies isolating individual component contributions. Removing the game-theoretic utility components (setting $\beta_3 = \beta_4 = 0$, eliminating congestion and redundancy terms) reduces hit ratio by 17% to 0.52 and increases redundancy index by 35% to 0.23. This confirms the game-theoretic components' crucial role in coordinating distributed decisions and managing externalities. Using only DDQN without PPO component reduces hit ratio by 12% to 0.55 due to less efficient exploration and difficulty maintaining diverse action distributions. Conversely, using only PPO without DDQN reduces performance by 9% to 0.57 from less sharp value estimates guiding exploitation. The hybrid architecture combining both provides complementary strengths: DDQN's efficient value-based exploitation and PPO's stable policy-based exploration [19]. Disabling potential-based reward shaping (removing the $\gamma[\Phi(s')-\Phi(s)]$ term from Equation 4) slows convergence by 40%, requiring 3500 episodes instead of 2500 to reach stable performance. Final performance also degrades slightly (hit ratio 0.61 vs 0.63) due to convergence to suboptimal local equilibria. This validates potential-based shaping's importance for guiding multi-agent learning toward globally efficient equilibria [41].

Removing the neighbor information exchange protocol (isolating routers without coordination) reduces hit ratio by 15% to 0.53 and increases redundancy to 0.28, nearly matching pure DRL-Cache. This demonstrates that explicit coordination mechanisms are essential for realizing game-theoretic benefits in distributed settings.

IV.6 ADAPTATION TO DYNAMIC WORKLOADS

To evaluate adaptation capabilities, we inject abrupt popularity shifts where Zipf parameter α changes from 0.7 to 1.2 at t=5000 seconds, creating flash crowd scenario where demand concentrates on fewer popular items. Figure 3 illustrates hit ratio evolution around this transition for different methods.

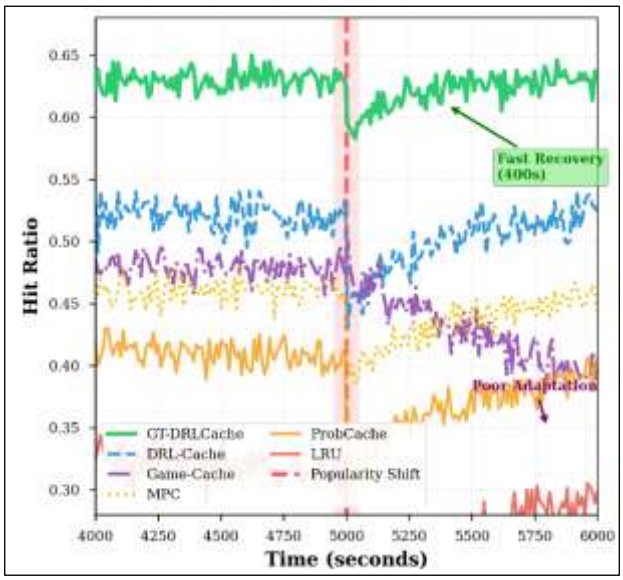


Figure 9: Hit Ratio Adaptation to Popularity Shift.
Source: Authors (2026).

GT-DRLCache exhibits rapid adaptation, with hit ratio temporarily dropping from 0.63 to 0.58 immediately after the shift but recovering to 0.62 within 400 seconds. The adaptive learning mechanisms quickly detect popularity changes through reward signals and adjust caching policies accordingly. The game-theoretic coordination ensures routers collectively adapt rather than acting independently, enabling faster recovery. Classical heuristics (LRU/LFU) show slower adaptation, requiring 1200+ seconds to reach new steady state performance. Their reactive nature means they must observe many cache misses before naturally adjusting cached content distribution. MPC performs better through prediction but still requires 800 seconds for Markov chain estimates to converge to new popularity distribution [6]. Game-Cache exhibits the poorest adaptation, with hit ratio dropping from 0.48 to 0.34 and never recovering due to its static equilibrium computation based on initial popularity distribution. This failure highlights fundamental limitations of non-adaptive game-theoretic approaches in dynamic environments. DRL-Cache adapts reasonably (600 second recovery) but suffers from coordination challenges during transition, with routers temporarily caching redundant content before independent learning realigns strategies.

IV.7 SCALABILITY ANALYSIS

We evaluate computational scalability by measuring training time and memory requirements across topologies of varying sizes. GT-DRLCache's training time scales approximately O(N log N) with network size N, growing from 3.2 hours for Abilene (12 nodes) to 14.5 hours for Barabási-Albert (50 nodes) on NVIDIA V100 GPU. This sublinear scaling stems from parallel training where router agents learn independently without centralized coordination bottleneck. Memory requirements scale linearly with network size, consuming 2.8GB for Abilene and 10.4GB for Barabási-Albert, primarily for experience replay buffers storing (state, action, reward) tuples. The per-router memory footprint remains constant around 200MB regardless of network size, enabling distributed deployment where each router maintains only local agent components [42].

Inference computational overhead is minimal, requiring only 0.3ms per caching decision for neural network forward passes. This enables real-time deployment even in high-traffic scenarios with thousands of requests per second. The trained policy networks require only 5MB storage per router, facilitating practical deployment on commodity hardware. Compared to centralized optimization approaches requiring $O(N^2)$ communication and computation, GT-DRLCache's distributed architecture provides substantial scalability advantages. The periodic coordination protocol incurs minimal overhead, exchanging only 50KB of neighbor information every 100 seconds, representing < 0.1% of typical link bandwidth.

IV.8 SENSITIVITY ANALYSIS

We analyze sensitivity to key hyperparameters to guide practical deployment. The discount factor $\gamma \in [0.9, 0.99]$ exhibits relatively stable performance, with optimal value $\gamma=0.95$ balancing short-term rewards and long-term planning. Very low values ($\gamma < 0.85$) cause myopic behavior ignoring future consequences, while very high values ($\gamma > 0.98$) overemphasize distant futures leading to slow learning. The utility weighting coefficients $\{\beta_k\}_{k=1}^4$ require careful tuning based on deployment objectives. We find $\beta_1=1.0$, $\beta_2=0.2$, $\beta_3=0.4$, $\beta_4=0.6$ provides good balance for general scenarios.

Increasing β_4 (redundancy penalty) reduces redundancy index but may slightly decrease hit ratio if set too high (> 1.0), while decreasing it increases redundancy without proportional hit ratio gains. Increasing β_3 (congestion penalty) improves traffic distribution but may increase latency if routers avoid caching popular content to prevent congestion. Network architecture choices show moderate sensitivity. Deeper networks (4-5 hidden layers) provide marginal improvements (1-2%) at significantly increased training time, while shallower networks (1-2 layers) underfit complex caching strategies reducing hit ratio by 5-8%. The selected 3-layer architecture balances expressiveness and efficiency. Learning rate exhibits typical sensitivity, with optimal value 0.0003 providing stable convergence. Higher rates (> 0.001) cause training instability and divergence, while lower rates (< 0.0001) converge slowly requiring 2-3x more training episodes. Adaptive optimizers like Adam provide robustness across reasonable learning rate ranges.

V. CONCLUSIONS

This paper presented **GT-DRLCache**, a unified framework synergistically integrating non-cooperative game theory with multi-agent deep reinforcement learning for intelligent distributed caching in Named Data Networking. By rigorously formulating the distributed caching problem as a finite stochastic potential game, we established theoretical foundations guaranteeing pure Nash equilibrium existence and enabling convergence analysis of the learning dynamics. The proposed hybrid Double-DQN/PPO architecture combines value-based and policy-based learning methods, providing both efficient exploitation through sharp Q-value estimates and stable exploration through smooth policy distributions. The potential-based reward shaping mechanism guides multi-agent learning toward globally efficient equilibria while preserving optimal policy invariance. The comprehensive utility function incorporating hit ratio improvement, replacement cost, link congestion, and redundancy overlap enables flexible optimization of multiple performance objectives.

Extensive simulation evaluation using **ndnSIM 2.8** on three diverse network topologies demonstrates GT-DRLCache's substantial performance advantages. Compared to state-of-the-art baseline methods, GT-DRLCache achieves 32–56% higher hit ratios, 28–44% latency reduction, and up to 50% upstream traffic savings. The framework exhibits robust performance across different network scales, connectivity patterns, and workload characteristics. Ablation studies confirm that both game-theoretic coordination and hybrid DRL components are essential for achieving superior performance. The practical implications of this work are significant for future information-centric networking deployments. The proposed framework provides a systematic approach to intelligent distributed caching that balances adaptive learning and strategic coordination. The computational efficiency enables real-time deployment on commodity hardware, while the distributed architecture provides natural scalability advantages. The theoretical foundations provide deployment confidence through convergence guarantees and equilibrium characterization.

Several limitations suggest directions for future research. First, the current evaluation focuses on medium-scale networks (12-50 nodes), and validation on larger production-scale topologies (1000+ nodes) would strengthen practical applicability claims. Second, the framework assumes cooperative routers following prescribed protocols, while real deployments may involve selfish operators requiring mechanism design for incentive alignment. Third, integration with adaptive routing strategies could enable joint optimization of caching and forwarding decisions. Future research directions include extending GT-DRLCache to hierarchical caching architectures common in content delivery networks, integrating with emerging 6G network paradigms incorporating edge computing and network slicing, developing privacy-preserving variants using federated learning for multi-operator scenarios, and incorporating online meta-learning for rapid adaptation to novel workload patterns. Additionally, theoretical analysis of convergence rates and approximation guarantees would strengthen the framework's foundations.

IV. AUTHOR'S CONTRIBUTION

Conceptualization: Seghier Djamal, Maaskri Moustafa, Goismi Mohamed and Debbab Mohamed.
Methodology: Seghier Djamal, Maaskri Moustafa, Goismi Mohamed and Debbab Mohamed.
Investigation: Seghier Djamal, Maaskri Moustafa, Goismi Mohamed and Debbab Mohamed.
Discussion of results: Seghier Djamal, Maaskri Moustafa, Goismi Mohamed and Debbab Mohamed.
Writing – Original Draft: Seghier Djamal, Maaskri Moustafa, Goismi Mohamed and Debbab Mohamed.
Writing – Review and Editing: Seghier Djamal, Maaskri Moustafa, Goismi Mohamed and Debbab Mohamed.
Resources: Seghier Djamal, Maaskri Moustafa, Goismi Mohamed and Debbab Mohamed.
Supervision: Seghier Djamal, Maaskri Moustafa, Goismi Mohamed and Debbab Mohamed.
Approval of the final text: Seghier Djamal, Maaskri Moustafa, Goismi Mohamed and Debbab Mohamed.

VI. REFERENCES

[1] L. Zhang et al., "Named Data Networking," ACM SIGCOMM Comput. Commun. Rev., vol. 44, no. 3, pp. 66-73, Jul. 2014. DOI: 10.1145/2656877.2656887

[2] S. Ullah, S. Hussain, I. Ali, H. Khattak, and S. Mastorakis, "Mitigating content poisoning attacks in named data networking: a survey of recent solutions, limitations, challenges and future research directions," Artificial Intelligence Review, vol. 58, no. 2, Dec. 2024. DOI: 10.1007/s10462-024-10994-x

[3] M. Amadeo, C. Campolo, A. Molinaro, and N. Mitton, "Named data networking: A natural design for data collection in wireless sensor networks," in Proc. IFIP Wireless Days, 2013, pp. 1-6. DOI: 10.1109/WD.2013.6686528

[4] I. Psaras, W. K. Chai, and G. Pavlou, "Probabilistic in-network caching for information-centric networks," in Proc. ACM ICN, 2012, pp. 55-60. DOI: 10.1145/2342488.2342501

[5] G. Rossini and D. Rossi, "Coupling caching and forwarding: Benefits, analysis, and implementation," in Proc. ACM ICN, 2014, pp. 127-136. DOI: 10.1145/2660129.2660153

[6] Y. Li, H. Xie, Y. Wang, and K. Poularakis, "MPC: Popularity-based caching strategy for content centric network," in Proc. IEEE ICC, 2013, pp. 3619-3623. DOI: 10.1109/ICC.2013.6655119

[7] Z. Liu, X. Zhang, and Y. Wang, "Proactive Caching With Distributed Deep Reinforcement Learning," IEEE Transactions on Parallel and Distributed Systems, vol. 35, no. 8, Aug. 2024. DOI: 10.1109/TPDS.2024.10540320

[8] L. Wang, Y. Chen, and Z. Xu, "Service caching with multi-agent reinforcement learning in cloud-edge collaboration computing," Peer-to-Peer Networking and Applications, Feb. 2025. DOI: 10.1007/s12083-025-01915-y

[9] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, Algorithmic Game Theory. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[10] S. Mastorakis, A. Afanasyev, I. Moiseenko, and L. Zhang, "ndnSIM 2: An updated NDN simulator for NS-3," NDN, Tech. Rep. NDN-0028, 2016.

[11] I. Psaras, R. G. Clegg, R. Landa, W. K. Chai, and G. Pavlou, "Modelling and evaluation of CCN-caching trees," in Proc. IFIP Netw., 2011, pp. 78-91. DOI: 10.1007/978-3-642-20757-0_7

[12] W. K. Chai, D. He, I. Psaras, and G. Pavlou, "Cache 'less for more' in information-centric networks," in Proc. IFIP Netw., 2012, pp. 27-40. DOI: 10.1007/978-3-642-30045-5_3

[13] J. Zhang, X. Li, Y. Wang, and Z. Chen, "XRL-SHAP-Cache: an explainable reinforcement learning approach for intelligent edge service caching in content delivery networks," Science China Information Sciences, Jun. 2024. DOI: 10.1007/s11432-023-3987-y

[14] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[15] Y. Li, "Deep reinforcement learning: An overview," arXiv:1701.07274, 2017.

[16] C. J. C. H. Watkins and P. Dayan, "Q-learning," Machine Learning, vol. 8, pp. 279-292, 1992. DOI: 10.1007/BF00992698

[17] V. Mnih et al., "Human-level control through deep reinforcement learning," Nature, vol. 518, pp. 529-533, Feb. 2015. DOI: 10.1038/nature14236

[18] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in Proc. AAAI, 2016, pp. 2094-2100.

[19] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv:1707.06347, 2017.

[20] Y. Chen, Z. Wang, and X. Liu, "Reinforcement Learning Based Approaches to Adaptive Context Caching in Distributed Context Management Systems," ACM Transactions on Internet of Things, 2024. DOI: 10.1145/3648571

[21] P. Hernandez-Leal, B. Kartal, and M. E. Taylor, "A survey and critique of multiagent deep reinforcement learning," Autonomous Agents Multi-Agent Systems, vol. 33, pp. 750-797, 2019. DOI: 10.1007/s10458-019-09421-1

[22] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in Proc. ICML, 1993, pp. 330-337.

[23] J. Nash, "Non-cooperative games," Annals of Mathematics, vol. 54, no. 2, pp. 286-295, 1951. DOI: 10.2307/1969529

[24] D. Monderer and L. S. Shapley, "Potential games," Games and Economic Behavior, vol. 14, no. 1, pp. 124-143, 1996. DOI: 10.1006/game.1996.0044

[25] R. W. Rosenthal, "A class of games possessing pure-strategy Nash equilibria," Int. J. Game Theory, vol. 2, pp. 65-67, 1973. DOI: 10.1007/BF01737559

[26] A. Araldo, G. Rossi, and D. Rossi, "Cost-aware caching: Caching more (costly items) for less (ISPs operational costs)," IEEE Trans. Parallel Distrib. Syst., vol. 27, no. 5, pp. 1316-1330, May 2016. DOI: 10.1109/TPDS.2015.2443811

[27] J. Lin, S. Huang, H. Zhang, X. Yang, and P. Zhao, "A novel coalition game based distributed cooperative content caching in mobile edge networks," Vehicular Communications, vol. 44, article 100689, Nov. 2023. DOI: 10.1016/j.vehcom.2023.100689

[28] A. Ioannou and S. Weber, "A distributed caching game with unknown content popularity," in Proc. IEEE INFOCOM, 2016, pp. 1-9. DOI: 10.1109/INFOCOM.2016.7524561

[29] H. Hu, G. Wen, and H. Wang, "Mean-field game theoretic edge caching in ultra-dense networks," IEEE Trans. Wireless Commun., vol. 17, no. 6, pp. 4020-4035, Jun. 2018. DOI: 10.1109/TWC.2018.2820713

[30] E. Baştuğ, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," IEEE Commun. Mag., vol. 52, no. 8, pp. 82-89, Aug. 2014. DOI: 10.1109/MCOM.2014.6871674

[31] Q. Wu, H. Zhang, Y. Li, and X. Wang, "Cooperative Edge Caching Based on Elastic Federated and Multi-Agent Deep Reinforcement Learning in Next-Generation Network," arXiv:2401.09886, Jun. 2024.

[32] Q. Wu, H. Zhang, Y. Li, and X. Wang, "Cooperative Edge Caching Based on Elastic Federated and Multi-Agent Deep Reinforcement Learning in Next-Generation Network," arXiv:2401.09886, Jun. 2024.

[33] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, "Mean field multi-agent reinforcement learning," in Proc. ICML, 2018, pp. 5571-5580.

[34] R. Mandala, "Analysis, Design, and Implementation of an Information Retrieval Approach for QoS Optimization in Information-Centric Networking Using Hybrid Naming Schemes," IEEE Access, vol. 13, pp. 166959-166972, 2025. DOI: 10.1109/ACCESS.2025.3610467

[35] M. Cha, H. Kwak, P. Rodriguez, Y. Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in Proc. ACM IMC, 2007, pp. 1-14. DOI: 10.1145/1298306.1298309

[36] V. Jacobson et al., "Networking named content," in Proc. ACM CoNEXT, 2009, pp. 1-12. DOI: 10.1145/1658939.1658941

[37] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," IEEE Trans. Commun., vol. 62, no. 10, pp. 3665-3677, Oct. 2014. DOI: 10.1109/TCOMM.2014.2352395

[38] Y. Sun, S. K. Fayaz, Y. Guo, V. Sekar, Y. Jin, M. A. Kaafar, and S. Uhlig, "Trace-driven analysis of ICN caching algorithms on video-on-demand workloads," in Proc. ACM CoNEXT, 2014, pp. 363-376. DOI: 10.1145/2674005.2674979

[39] G. Zhang, Y. Li, and T. Lin, "Caching in information centric networking: A survey," Comput. Netw., vol. 57, no. 16, pp. 3128-3141, Nov. 2013. DOI: 10.1016/j.comnet.2013.07.007

[40] K. Cho, M. Lee, K. Park, T. T. Kwon, Y. Choi, and S. Pack, "WAVE: Popularity-based and collaborative in-network caching for content-oriented networks," in Proc. IEEE INFOCOM Workshops, 2012, pp. 316-321. DOI: 10.1109/INFCOMW.2012.6193512

[41] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in Proc. ICML, 1999, pp. 278-287.

[42] E. Bastug, M. Bennis, M. Médard, and M. Debbah, "Toward interconnected virtual reality: Opportunities, challenges, and enablers," IEEE Commun. Mag., vol. 55, no. 6, pp. 110-117, Jun. 2017. DOI: 10.1109/MCOM.2017.1601089

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 551-558. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.1509

RESEARCH ARTICLE OPEN ACCESS

INTEGRATED MCDM MODEL FOR EVALUATION TWO FILE PREQUALIFICATION TENDER AT PROCUREMENT BUREAU IN INDONESIA

Citra Amalia[1], Afiza Ismail[2], Emma Nuraihan Mior Ibrahim[3], Shuhaida Mohamed Shuhidan[4] and Norzehan Sakamat[5]

[1,2,3]Universiti Teknologi MARA-College of Computing, Informatics, and Mathematics. Shah Alam,Malaysia
[4]Universiti Teknologi MARA-Centre of Foundation Studies. Dengkil, Malaysia.
[5]Universiti Teknologi Petronsl-Computer and information Science. Perak, Malaysia.

[1]https://orcid.org/0000-0003-2779-2776, [2]https://orcid.org/0009-0005-8927-0713, [3]https://orcid.org/0000-0001-8916-2978, [4]https://orcid.org/0000-0002-6317-4724, [5]https://orcid.org/0000-0002-2170-8249

Email: citra.chicit@gmail.com, afiza025@uitm.edu.my, emmanuraihan@uitm.edu.my, shuhaida.shuhidan@utp.edu.my, norzehan012@uitm.edu.my

ARTICLE INFO

Article History
Received: December 29, 2024
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Tender,
LPSE,
MCDM,
ANP,
WASPAS.

ABSTRACT

Tenders in Indonesia are managed by the government, with each province having a goods and services procurement bureau that serves as the selection committee known as POKJA (*Kelompok Kerja*). The government uses LPSE for procurement, with tender evaluations involving document verification. The lowest bid prices and complete documents from three companies are called to attend a document verification evaluation at the Office of the Procurement Bureau. The verification evaluation process is typically manual and based on the working group's opinion, with no system involved. This results in unrealistic execution times, high qualification requirements, short auction processes, and poor coordination between the Commitment Making Officer and POKJA. To address these issues, a systematic and transparent auction evaluation procedure with clear criteria is needed, with multi-criteria decision making being the most appropriate method after analysis. The study on tendering and selection integration of Multi-Criteria Decision Making (MCDM) using ANP and WASPAS methods, which are better for complex scenarios and provide accurate, error-reducing results, making the study more mathematically straightforward. The study aims to aid the government in implementing transparant e-procurement through a model developed by procurement bureau. This model will enable a more transparent tender process, resulting in clearer contractor assessment results.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Tenders in Indonesia have a strong legal basis and are included in the JDIH LKPP (Legal Documentation and Information Network of the Government Procurement Policy Agency). For the construction sector, the regulations are governed by the Regulation of the Government Procurement Policy Agency No. 20 of 2021 concerning Guidelines for the Implementation of Government Procurement of Goods/Services through Providers. Procurement of Goods/Services is an activity that starts from the identification of needs to the handover of the work results. The guidelines for the implementation of Goods/Services Procurement through Providers include the preparation activities for Goods/Services Procurement through Providers, the preparation for the selection of Providers, the implementation of the selection of Providers, the execution of the Contract, and the handover of the work results. Before the procurement implementation, a Needs Analysis and Evaluation, as well as Procurement Planning, are conducted. The implementation of Goods/Services Procurement through Providers is a continuation of the procurement planning that has been carried out by the Budget User. The procurement officer who conducts the tender selection is the POKJA/Working Group.

The Selection Working Group determines the method of selecting Providers by considering the type of Construction Work and Construction Consultancy Services, Technical Specifications/Terms of Reference, and the complexity of the work, Budget Ceiling/Estimated Price, contract draft, market analysis results, and/or consolidation results. The Selection Committee implements the method of selecting Construction Service Providers, including: Direct Appointment, Quick Tender, Tender. Tender is used when E-Purchasing, Direct Procurement, Direct Appointment, and Quick Tender cannot be utilized. In a tender, there is a qualification method. Qualification is an evaluation of competence, business capability, and compliance with requirements as a Provider. Qualification is conducted through post-qualification or pre-qualification. This research focuses on pre-qualification tenders. The qualification evaluation in pre-qualification uses a knockout system for Construction Work Providers and a direct appointment process, or a weighted system with a threshold for Construction Consultancy Service Providers.

The pre-qualification results for Construction Work Tenders require at least 3 (three) participants to pass the qualification, for the Selection of Construction Consultancy Services, at least 3 (three) and at most 7 (seven) participants to pass the qualification, or for Direct Appointments, the pre-qualification results meet or do not meet the qualification requirements. If the participants who pass the qualification for Tenders/Selections are less than 3 (three), the pre-qualification is declared failed and a re-pre-qualification is conducted. [1]. The government, through the LKPP as the highest institution in Indonesia that conducts procurement practices, established the LPSE or Electronic Auction Institution. Based on Presidential Regulation Number 54/2010 jo. Presidential Regulation Number 70/2012, the LKPP was mandated to develop the Electronic Procurement System (SPSE) [2]. The version of SPSE currently in use is SPSE v4.5. The POKJA announces the tender until the winner is announced on the LPSE website.

One of the methods discussed in this research is the tender with the 2-file prequalification method, where participants will be called to attend the evaluation stage of the bid document verification, which consists of administrative, technical, and price documents. Subsequently, the Selection Committee conducts qualification verification for participants who meet the qualification requirements after the qualification evaluation is completed. The invited companies are 3 qualified companies to verify and explain all documents. The qualification verification evaluation process consists of 3 main criteria: administration, technical, and price. The first stage is the Opening of the Administrative and Technical Offer Documents (file I), where the Selection Committee evaluates the Offer Documents based on the data inputted and uploaded in SPSE. Next, the Selection Committee opens the price offer documents (file II). The price offer documents (file II) of participants who do not pass the administrative and technical evaluations are not opened. However, if they meet the requirements, the price offer documents (file II) will be opened.

The evaluation of the qualification opening is conducted privately between the POKJA and the selected Company, which will be called according to the schedule. The assessment is carried out manually by the POKJA, with the scoring following the POKJA's Decision. To avoid fraud, a system was created in the qualification proof evaluation process to make the tender process transparent, as tender participants can understand the evaluation processes of administration, technical, price, and qualification, thereby minimizing fraud in the tender process. [3]. During the data collection through observation, the researcher found inconsistencies, such as one company lacking documents in the expert personnel section, yet the committee still deemed the documents sufficient. According to the guidelines, this company should not meet the requirements and should be disqualified. However, due to the absence of applicable criteria in this case, it was considered qualified depending on whether the company could convince the committee or not. Such practices are highly discouraged, therefore a precise, accurate, and transparent evaluation system is needed in the qualification verification process. After conducting an analysis to determine the appropriate system method in this case, it was decided to use Multi Criteria Decision Making. (MCDM).

Multi-criteria decision-making (MCDM) is one of the biggest decision-making issues which seeks to discover the best alternative by considering more than one factor in the selection process. MCDM provides several tools and strategies that may be employed in a variety of sectors. This post intends to offer a survey of the MCDM Model and integration methods.[4]. Creating a complete model for tender evaluation using MCDM techniques entails including a variety of factors to enable well-informed selections [5]. The primary goal of the MCDM technique is to choose the best choice from a group of options in the face of several competing criteria. However, there is a gap in the integration of uncertainty and sensitivity analyses. Future study should concentrate on sensitivity analysis to determine decision-making resilience to modifications in criterion weights, inputs, and assumptions. Time dynamics and temporal concerns should be taken into account throughout the MCDM process with ANP and WASPAS. [6-9] .Hybrid decision-making models that include several MCDM strategies should be investigated, including synergies with other approaches such as TOPSIS, AHP, and ELECTRE. [10].

II. THEORETICAL REFERENCE

II.1 EVALUATION TWO FILE PREQUALIFICATION TENDER

Prequalification is the process of evaluating qualifications conducted before the submission of bids. Prequalification is conducted during the implementation of Construction Work Tenders for complex procurement. The result of the prequalification for Construction Work Tenders is at least 3 (three) participants who pass the qualification. The Two File Method is used for the Procurement of Other Goods/Services that require a technical assessment first, namely the Value System Evaluation Method, the Life Cycle Cost Evaluation Method, and the Lowest Price Evaluation Method that uses threshold weighting. The stages of the 2-file Prequalification process include:

Invitation to participants who pass the prequalification, downloading of selection documents, provision of explanations, submission of Bid Documents, opening of administrative and technical Bid Documents (file I), administrative evaluation, technical evaluation for those who pass the administrative evaluation, announcement of participants who pass the administrative and technical evaluations (file I), opening of Bid Documents (file II) for those who pass the technical evaluation, price evaluation, determination and announcement of the winner, objection period, and the Selection Working Group's report to the Commitment Making Officer (PPK). [1] The implementation of prequalification begins with the Announcement of Prequalification, Registration and Downloading of Qualification Documents, Provision of Qualification Explanations (if necessary), Submission of Qualification Documents, Evaluation of Qualification Documents, Qualification Proof, Determination of Qualification Results, Announcement of Qualification Results,

Qualification Objection, Follow-up on Failed Prequalification. Two Files is a method of submitting bid documents where the administrative and technical requirements are included in one (1) file, while the bid price is included in another file and uploaded separately at the same time. The form of the Administrative and Technical Offer Document (File I) is as follows. For the Technical evaluation (File I), the Administrative Offer Document includes the Offer Letter (according to SPSE) and the bid guarantee (if required). Furthermore, for the Technical Offer Document, there are the Work Implementation Method, List of Main Equipment, List of Managerial Personnel, List of Subcontracted Work Sections (if required), Construction Safety Plan Form (RKK); and other required documents (if required). The form of the Price Offer Document (File II) is the Offer Price according to the Offer Letter and the Quantity and Price List (for unit price work sections) as well as the Output and Price List (for lump sum work sections). [1]

II.2 PROCUREMENT BUREAU

The Bureau of Goods and Services Procurement (BPBJ) has the task of assisting the Assistant for Economic and Development Affairs in preparing regional policy formulation, coordinating the implementation of Regional Apparatus tasks, coordinating the implementation of Regional Apparatus tasks in the field of goods and services procurement management, managing electronic procurement services, and providing guidance and advocacy for goods and services procurement. The Bureau of Procurement of Goods and Services consists of the Goods and Services Procurement Management Section, the Electronic Procurement Services Management Section, and the Goods and Services Procurement Guidance and Advocacy Section [11]. The Goods and Services Procurement Management Division has the task of preparing materials for regional policy formulation, coordinating regional policy formulation, coordinating the implementation of goods and services procurement, monitoring and evaluating the implementation of regional policies in the field of goods and services procurement strategy management, implementing goods and services procurement, and monitoring and evaluating goods and services procurement.

The Electronic Procurement Services Management Section is tasked with preparing materials for regional policy formulation, coordinating regional policy formulation, coordinating the implementation of electronic procurement services management, monitoring and evaluating the implementation of regional policies in the field of electronic procurement system management, developing information systems, and managing procurement information for goods and services. The Procurement and Advocacy Division for Goods and Services has the task of preparing materials for the formulation of regional policies, coordinating the implementation of training and advocacy for the procurement of goods and services, monitoring and evaluating the implementation of regional policies in the field of human resource development for the procurement of goods and services, institutional development for the procurement of goods and services, and providing assistance, consultation, and/or technical guidance for the procurement of goods and services.[11]. BPBJ formed a Working Group (POKJA) team for all tenders in each region.

The Selection Working Group can adjust the Tender Documents as needed as long as it does not conflict with the laws and regulations. The identity of the selection committee as stated in the Election Data Sheet (LDP). Based on Presidential Regulation No. 16 of 2018 Article 1 number 12, the Selection Working Group (POKJA) hereinafter referred to as the Selection Working Group is the human resource designated by the UKPBJ leadership to manage the selection of Providers. According to Presidential Regulation No. 16 of 2018 Article 13, the Selection Working Group in the Procurement of Goods/Services as referred to in Article 8 letter e has the task of preparing and implementing the selection of Providers, preparing and implementing the selection of Providers for the electronic catalog, and determining the winners of the selection/Providers for the tender and selection methods. The Selection Working Group consists of 3 (three) members. In cases where the complexity of the Provider selection warrants it, the members of the Selection Working Group may be increased as long as the number is odd. The Selection Committee can be assisted by a team or experts. [12]

II.3 LPSE

This LPSE system is a system developed by the Government Goods/Services Procurement Policy Agency. (LKPP). The government, which is a non-departmental government agency under and responsible to the President, was established based on Presidential Regulation Number 106 of 2007. LKPP is the only government agency responsible for developing and formulating government procurement policies for goods/services. In carrying out its duties and functions, LKPP is coordinated by the Minister of National Development Planning. And now, the conveniences in participating in government procurement activities have been developed by LKPP (the government procurement policy agency) through the implementation of the e-procurement and e-catalog systems, which are integrated into one system called LPSE (electronic procurement service). Once registered in the LPSE system, participants are ensured to be able to engage in procurement activities throughout Indonesia.[13]

II.4 MULTI CRITERIA DECISION MAKING (MCDM)

Since the 1970s, the literature on multi-criteria decision making has grown dramatically, and a wide range of decision-making approaches and tools have been created for a variety of choice issues.[14]. The research study found that sustainable supplier selection by MCDM had considerable increase in 2019. Researchers mostly communicated their scientific discoveries through articles, accounting for 81% of the publications, followed by conference papers (14%), and book chapters (2.5%); [15]. MCDM is a decision-making approach or process that considers various options based on a number of criteria. MCDM is a part of the scientific field of operations research.

The language's emphasis mixes qualitative and quantitative components, multi-criteria that sometimes contradict each other on the achievement of performance, therefore the value of the criterion is balanced. [16]. The operationalization of MCDM techniques is generally classified into 2 categories, namely Multi Attribute Decision Making (MADM) based on discrete patterns and Multi Objective Decision Making based on continuous patterns (MODM) as well as a combination of both [17]. Multi criteria decision making (MCDM) is a well-established branch of operations research, with core ideas dating back to the 18th century. Methods like ELECTRE, AHP, TOPSIS, and PROMETHEE emerged in the latter half of the 20th century[18].

Alinezhad and Khalili highlight that only a few methodologies, like MOORA (2006) and WASPAS (2012), are widely used in business or engineering due to factors like overabundance of techniques, poor knowledge diffusion, and timescale.

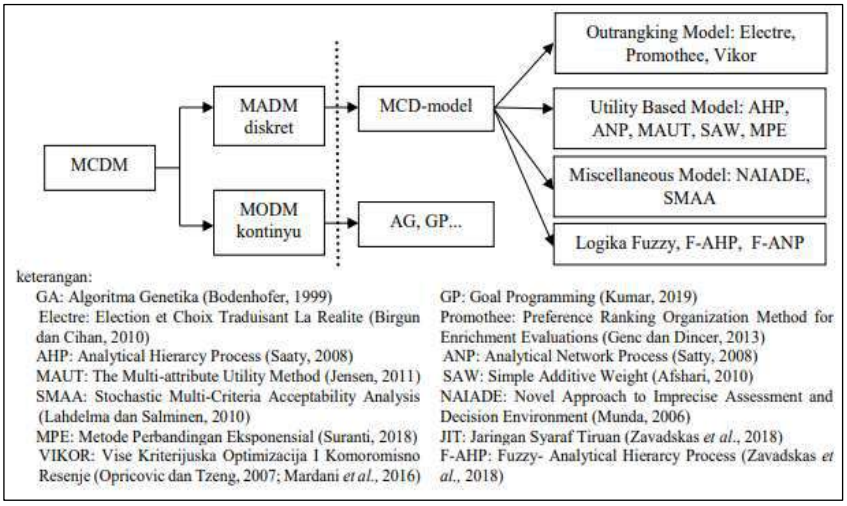


Figure 1: Structure MCDM and MCD-Model.
Source: [16].

III. MATERIALS AND METHODS

The focus of this research is solely on e-procurement tenders in the construction industry in Indonesia that use LPSE as a medium for conducting e-tenders. The object of the research is the Procurement Bureau of Goods and Services of Central Kalimantan Province. Data collected based on sources of observation, interviews, focus group discussions, and government documents related to construction tenders based on legal foundations. For the process of evaluating construction tenders based on literature studies and the experience of the tender committee/POKJA. In this study, the researcher uses two MCDM techniques to develop a multi-criteria decision support model in construction tenders to select the winning contractor. based on the literature review, 2 MCDM methods were chosen in this study, namely Analysis Network Process (ANP) and The Weighted Aggregated Sum Product Assessment (WASPAS). The research methodology framework is a crucial aspect of a research project that may clearly describe the workflow from beginning to end. As stated in the figure below, this research will be conducted in five phases, each of which will consist of 3-6 steps.

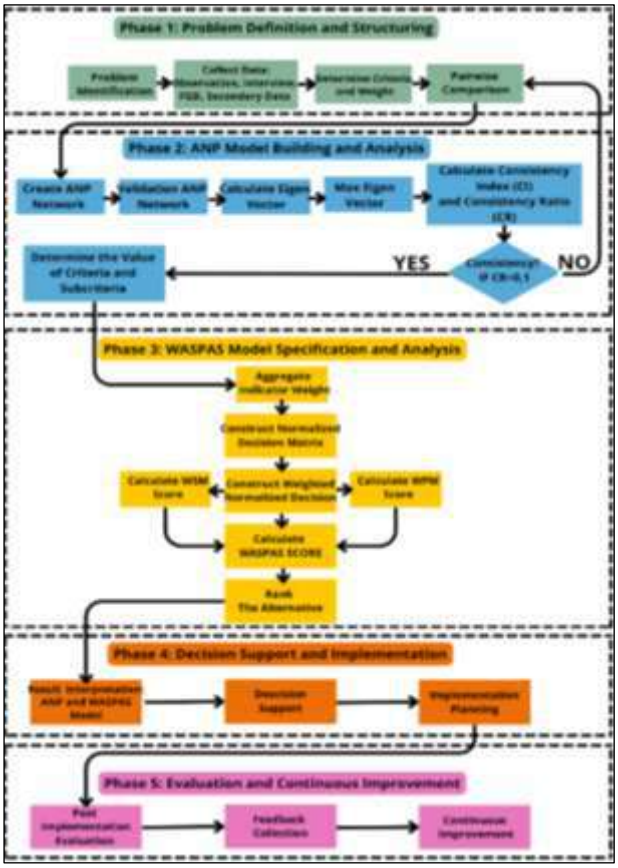


Figure 2: Research Framework.
Source: Authors, (2026).

Phase 1 to determine the criteria in the selection of contractor for a tender. Phase 2 and 3 to propose MCDM model for selection of contractor in tender evaluation. Phase 4 and 5 to evaluate the accurancy, precision, and specificity of the proposed MCDM Model in reducing biases and increase decision consistency.

IV. RESULTS AND DISCUSSIONS

After conducting analysis and data collection through observation, interviews, focus group discussions, and secondary data from BPBJ Central Kalimantan, three main criteria were identified in the implementation process of qualification proof evaluation for the two-file prequalification tender type, namely Administrative, Technical, and Price Criteria. Each of these three criteria has sub-criteria within them, as shown in the table below.

Table 1: Table Criteria & SubCriteria Two File Prequalification.

Criteria	Sub Criteria
Administrative	Tender Document (Tehnical & Price Document); Tender Guarantee; Administrative Data; Legal Basis for Establishing a Business Entity; Business Entity Management; Business License; Business Entity Certificate; Other Certificates (if required); Financial Data (shareholding structure and taxes); Company Experience Data (last 15 years); Company Experience Data (last 3 years); Ongoing Project Data
Technical	Method of Work Execution; Main Equipment; Managerial Personnel; List of subcontracted work sections in the form of types of work subcontracted according to the SSKK provisions (if required); Construction Safety Plan Document
Price	Evaluation of Offer Price; Evaluating the completeness of the price offer documents and arithmetic corrections; The total corrected offer price compared to the HPS value; Evaluation of Price Fairness Considering price preferences for the use of domestic production (if it meets the requirements for the application of price preferences)

Source: Author (2026).

The tender evaluation is conducted using a lowest price system with a threshold. The committee evaluates the tender Documents based on the data inputted and uploaded in SPSE, except for the evaluation of the tender Guarantee, which is conducted based on the submitted tender Guarantee document. The Selection Committee conducts the evaluation of the first file offers, which includes: administrative evaluation and technical evaluation. The Selection Committee inputs the results of the first file offer document evaluation into the SPSE and displays the results of the first file evaluation through the announcement menu or other information upload menu on the SPSE.

Subsequently, the Selection Committee opens the second file offers with the following provisions: The second file offer documents of participants who do not pass the administrative and technical evaluations are not opened; The Selection Committee is not allowed to disqualify offers at the time of opening the second file offer documents, unless the second file offers cannot be opened (decrypted) based on information from the LPSE; After the second file offers are opened, the Selection Committee conducts a price evaluation.

Administrative evaluation using a knockout system. Technical evaluation has a score for each of its subcriteria, namely Method of Work Execution (44%); Main Equipment (24%); Managerial Personnel(20%); List of subcontracted work sections in the form of types of work subcontracted according to the SSKK provisions (if required)(7%); Construction Safety Plan Document(5%). Price evaluation has a score of 1-100 for each sub-criterion. Prequalification tender two files using the merit point system, with 80% for technical criteria and 20% for price criteria. Furthermore phase 2, ANP Model and analysis with 6 Step that is Create Network, Validation ANP Network, Calculate Eigen Vector, Max Eigen Vector, Calculate Consistency Index and Consistency Ratio, and Determine the Value of Criteria and Subcriteria.

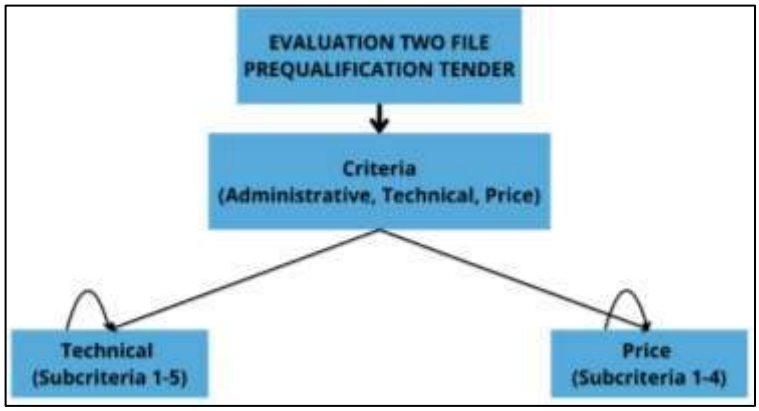


Figure 3: ANP Network.
Source: Author (2026).

The Analytic Network Process (ANP) is a decision-making method developed by Thomas L. Saaty, which enhances the Analytic Hierarchy Process (AHP) by modeling relationships among elements in a network structure. It is widely used in construction research for condition assessment, technology evaluation, and safety assessment, assisting in calculating the importance of criteria and facilitating systematic decision-making in complex scenarios [19-21]. Calculate Eigen Vector, The eigenvalue of a matrix is determined by summing the values of each column, dividing each column cell by the total column, and summing the values of each row and dividing by n. This process involves dividing by n and dividing by n.

Eigen Vector

$$X = \frac{\sum W_{i,j} / \sum W_j}{\text{n}} \quad (1)$$

Where X : *eigenvector;* Wij : column cell value in one row (i,j = 1….n); ∑Wj : total column; n : number of matrices being compared.

Max Eigen Vector : λ_{maks} which is the largest value of the eigenvector.

Consistency ratios verify accuracy of assessments for ANP calculations, with a CR value of less than 0.1 indicating consistency in the comparison matrix.

Consistency Index and Consistency Ratio

$$CI = \frac{\lambda_{maks} - n}{n - 1} \quad (2)$$

In a consistency matrix, practically λ_max=n, whereas in a matrix not every variation of will bring a change in the value of λ_max. The deviation of λ_max from n is a Consistency Index (CI) parameter. Description: CI= Consistency Index; λ_max= largest eigenvalue ; n = number of elements being compared

$$R = \frac{CI}{RI} \quad (3)$$

Where: CR = Consistency Ratio; CI = Consistency Index; RI = Random Index.

Phase 3 involves analyzing the WASPAS model for contractor selection in tender evaluation. The process involves aggregating indicator weight, constructing a normalized decision matrix, calculating WSM, WPM, Waspas scores, and ranking alternatives. This structured approach incorporates multiple criteria and stakeholder preferences, providing a structured approach to contractor selection.

Aggregate Indicator Weight

$$=\sum_{i=1}^{n}(Normalized\ Score_i \times Normalized\ Weight_i \quad (4)$$

The aggregate indicator weight is a crucial factor in evaluating the overall performance of alternatives, aiding decision-makers in prioritizing and comparing options based on a thorough review of numerous criteria. Construct Normalized Decision Matrix : Normalized decision matrix with weighted scores reflecting the total performance of each alternative taking into account the relevance of each criterion. This matrix can be used to rank alternatives in a decision-making problem.

Normalized Decision Matrix

$$\bar{X}ij = \frac{min_i(X_{ij})}{X_{ij}} \quad (5)$$

Where: Xij: is the normalized value for a specific tuple (alternative, criterion); Xij is the original value to be normalized; mini(Xij) represents the smallest value within the set of values for a specific monotonic cost criterion across all alternatives, the index i ranges from 1 to m, representing the number of alternatives.

Construct Weight Normalized Decision: Constructing the Decision Matrix MCDM challenges are characterized by sets of m choices and n criteria. As a result, an m-by-n matrix is generated with known values of the n criterion for each of the m possibilities.

Calculate WSM Score : The WSM score is calculated by adding the normalized scores and their appropriate weights for each option. This approach gives an easy way to assess and rank alternatives based on their overall performance when numerous criteria are considered. Adjust the weights and normalization procedures to meet the unique needs of the decision-making problem.

Calculate WPM Score: The WPM score is calculated by multiplying the product of the normalized scores with the weights assigned to each alternative. This approach gives an alternate method for evaluating and ranking alternatives based on their overall performance when numerous criteria are considered.

Calculate WASPAS Score: The WPM score is calculated by multiplying the product of the normalized scores with the weights assigned to each alternative. This approach gives an alternate method for evaluating and ranking alternatives based on their overall performance when numerous criteria are considered

Rank the Alternative: Ranking the results of the above calculations.

The integration of multiple Multi-Criteria Decision-Making (MCDM) methods, like F-ANP and F-WASPAS [10], can enhance the decision-making process. This approach considers a wide range of criteria and perspectives, reducing bias and ensuring a balanced, unbiased process. It also provides additional insights and validation, mitigates risks, enhances transparency, and increases stakeholder confidence in the chosen course of action, ultimately improving decision quality and resource allocation in complex scenarios.



Figure 4: Result Rank Evaluation from LPSE.
Source: [13].

V. CONCLUSIONS

The integration of the Analytic Network Process (ANP) and Weighted Aggregated Sum Product Assessment (WASPAS) methods involves combining the strengths of both techniques to enhance the decision-making process in Multi-Criteria Decision Making (MCDM) scenarios. The integration of ANP and WASPAS involves using the ANP framework to determine the weights and priorities of criteria and alternatives within a decision network. These values are then used in the WASPAS method to evaluate and rank alternatives based on weighted attribute values. The ANP-derived weights ensure the evaluation of alternatives reflects the relative importance of criteria. This allows decision-makers to analyze complex decision structures and prioritize alternatives based on weighted criteria values. The integration of ANP and WASPAS provides a robust approach to MCDM, combining the analytical depth of ANP with the evaluative power of WASPAS. This comprehensive approach considers structural relationships and dependencies between criteria (ANP) and the aggregation of preferences across multiple criteria (WASPAS), enabling more informed and structured decision-making in complex systems.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Citra Amalia, Afiza Ismail, Emma Nuraihan Mior Ibrahim, Shuhaida Mohamed Shuhidan and Norzehan Sakamat.
Methodology: Citra Amalia, Afiza Ismail.
Investigation: Citra Amalia, Afiza Ismail.
Discussion of results: Citra Amalia, Afiza Ismail, Emma Nuraihan Mior Ibrahim, Shuhaida Mohamed Shuhidan and Norzehan Sakamat.
Writing – Original Draft: Citra Amalia.
Writing – Review and Editing: Citra Amalia, Afiza Ismail.
Resources: Afiza Ismail.
Supervision: Afiza Ismail, Shuhaida Mohamed Shuhidan and Norzehan Sakamat.
Approval of the final text: Citra Amalia, Afiza Ismail, Emma Nuraihan Mior Ibrahim, Shuhaida Mohamed Shuhidan and Norzehan Sakamat.

VII. ACKNOWLEDGMENTS

First of all, I would like to thank the College of Computing, Informatics, and Mathematics at Universiti Teknologi MARA for all the assistance provided in this research.

VIII. REFERENCES

[1] J. LKPP, "Peraturan Lembaga Nomor 12 Tahun 2021," JDIH LKPP. Accessed: Dec. 27, 2024. [Online]. Available: https://jdih.lkpp.go.id/regulation/peraturan-lkpp/peraturan-lkpp-nomor-12-tahun-2021

[2] S. Nadjamuddin and M. Nawawi, "EVALUASI PENERAPAN LPSE DALAM PERSPEKTIF GOOD GOVERNANCE," *Jurnal Informatics*, vol. 1, pp. 2338–9834, 2020, [Online]. Available: http://www.lkpp.go.id/].

[3] Y. Salupra, "EVALUATION OF E-TENDERING IMPLEMENTATION AT REGIONAL PROCUREMENT SERVICE UNIT OF MINISTRY OF FINANCE IN SOUTH SULAWESI PROVINCE," *Jurnal Administrasi Negara*, vol. 26, no. 1, pp. 46–67, 2020.

[4] H. Taherdoost and M. Madanchian, "Multi-Criteria Decision Making (MCDM) Methods and Concepts," *Encyclopedia*, vol. 3, no. 1, pp. 77–87, Jan. 2023, doi: 10.3390/encyclopedia3010006.

[5] P. Ariya and K. Puritat, "The Multi-Criteria Decision Model using ANP approach for selection ERP Software of SMEs in the Upper North Region of Thailand," *2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering, ECTI DAMT and NCON 2020*, pp. 206–210, 2020, doi: 10.1109/ECTIDAMTNCON48261.2020.9090749.

[6]S. Chakraborty, P. Chattererjee, and P. P. Das, *Multi-Criteria Decision-Making Methods in Manufacturing Environments Models and Applications*. CRC Press, 2023. Accessed: Jan. 03, 2024. [Online]. Available: https://books.google.com.my/books?id=gffOEAAAQBAJ&dq=AHP+ANP++BOOK&lr=&hl=id&source=gbs_navlinks_s

[7] E. Yunaeti Anggraeni and S. Hartati, "Implementasi Metode Waspas (Weight Aggregated Sum Product Assesment) Dalam Menentukan Ruko Yang Strategis," *Jurnal Teknologi Informasi*, vol. 3, 2021.

[8] M. I. Alfarisi and M. Safii, "Penerapan Metode WASPAS Dalam Menentukan Peminatan Jurusan Pada Penerimaan Mahasiswa Baru," *Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD*, vol. 6, no. 1, pp. 274–280, 2023, [Online]. Available: https://ojs.trigunadharma.ac.id/index.php/jsk/index

[9] C. Lukita, C. Nas, and W. Ilham, "Analisis Pengambilan Keputusan Penentuan Prioritas Utama Dalam Peningkatan Kualitas Mata Pelajaran Dengan Menggunakan Metode Perbandingan WASPAS dan MOORA," *Jurnal Nasional Teknologi dan Sistem Informasi* , 2019.

[10] M. Saghari, M. Ehsanifar, and A. Babaei, "Identification and Prioritization of Optimal Places for Locating Urban Hospitals by Combin-ing the F-ANP, F-WASPAS Methods, and GIS Output in the ArcGIS Environment: Case Study: Arak City," *Theory of Approximation and Applications*, vol. 14, no. 2, pp. 13–34, 2020, doi: 20.1001.1.25382217.2020.14.2.10.4.

[11] UKPBJ, "Tugas Pokok dan Fungsi Biro Pengadaan Barang dan Jasa Sekretariat Daerah Provinsi Kalimantan Tengah," BIRO PBJ KALTENG. Accessed: Dec. 28, 2024. [Online]. Available: https://biropbj.kalteng.go.id/dokumen/detail/tugas-pokok-dan-fungsi-biro-pengadaan-barang-dan-jasa-sekretariat-daerah-provinsi-kalimantan-tengah

[12] ukpbj admin, "Dasar Hukum Kelompok Kerja Pemilihan (POKJA Pemilihan)," Bagian Pengadaan Barang/Jasa Setda Kota Depok-Pemerintah Kota Depok. Accessed: Dec. 28, 2024. [Online]. Available: https://ukpbj.depok.go.id/?p=106

[13] J. Pangaribuan, Safuan, and Musa, "PENERAPAN LPSE (LAYANAN PENGADAAN SECARA ELEKTRONIK) UNTUK KEMUDAHAN KEGIATAN PENGADAAN BARANG DAN JASA," *Syntax Literate: Jurnal Ilmiah Indonesia*, vol. 7, no. 2, Feb. 2022.

[14] J. van der Meer, A. Hartmann, A. van der Horst, and G. Dewulf, "Multi-criteria decision analysis and quality of design decisions in infrastructure tenders: a contractor's perspective," *Construction Management and Economics*, vol. 38, no. 2, pp. 172–188, Feb. 2020, doi: 10.1080/01446193.2019.1577559.

[15] K. Khulud, I. Masudin, F. Zulfikarijah, D. P. Restuputri, and A. Haris, "Sustainable Supplier Selection through Multi-Criteria Decision Making (MCDM) Approach: A Bibliometric Analysis," Dec. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/logistics7040096.

[16] R. Jaya, E. Fitria, Y. Yusriana, and R. Ardiansyah, "MULTI-CRITERIA DECISION MAKING (MCDM) IMPLEMENTATION ON AGROINDUSTRIAL: A LITERATURE REVIEW," *Jurnal Teknologi Industri Pertanian*, pp. 234–343, Sep. 2020, doi: 10.24961/j.tek.ind.pert.2020.30.2.234.

[17] R. Jaya, E. Fitria, Yusriana, and R. Ardiansyah, "Implementasi Multi Criteria Decision Making (Mcdm) Pada Agroindustri: Suatu Telaah Literatur," *Jurnal Teknologi Industri Pertanian*, pp. 234–343, Sep. 2020, doi: 10.24961/j.tek.ind.pert.2020.30.2.234.

[18] K. Urošević, Z. Gligorić, I. Miljanović, C. Beljić, and M. Gligorić, "Novel methods in multiple criteria decision-making process (Mcrat and raps)—application in the mining industry," *Mathematics*, vol. 9, no. 16, Aug. 2021, doi: 10.3390/math9161980.

[19] A. Septian, "Subcontractor Selection Review Industry XYZ With Analytic Network Process (ANP) Method," *IPTEK Journal of Proceeding Series No. 6 International Seminar on Science Technology (ISST)*, vol. 6, pp. 558–562, Jul. 2020.

[20] M. Assaf, M. Hussein, S. Abdelkhalek, and T. Zayed, "A Multi-Criteria Decision-Making Model for Selecting the Best Project Delivery Systems for Offsite Construction Projects," *Buildings*, vol. 13, no. 2, Feb. 2023, doi: 10.3390/buildings13020571.

[21] M. R. Ismail, M. Sun, and G. Bowles, "A risk-oriented tender evaluation system for construction projects in Malaysia," *Engineering, Construction and Architectural Management*, pp. 1–33, Jul. 2021, [Online]. Available: http://mc.manuscriptcentral.com/ecaam

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 559-566. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2567

RESEARCH ARTICLE OPEN ACCESS

A THREE-TIER INTEGRATED QUEUE MANAGEMENT SYSTEM: OPTIMIZED SERVICE DELIVERY FOR LAND TRANSPORTATION CLIENTS

Ines G. Falcon*[1], Kenneth Jay L. Dugaria[2] and Danna C. Remojo[3]

[1, 2, 3]Department of Information Information Technology, Southern Leyte State University, 6611 Southern Leyte, Philippines.

[1]http://orcid.org/0000-0003-0660-5302, [2]http://orcid.org/0009-0009-8292-9692, [3]http://orcid.org/0009-0004-2236-2915

Email: *ifalcon@southernleytestateu.edu.ph, kdugaria@southernleytestateu.edu.ph, dremojo@southernleytestateu.edu.ph

ARTICLE INFO

Article History
Received: August 26, 2025
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
automation,
integrated queue management,
lining up,
optimized service delivery.

ABSTRACT

Optimization of client satisfaction and faster services are the main goals of any service-oriented government organization. However, the land transportation agency in Eastern Visayas Region of the Philippines experience challenges in the management of its flow and processes, as the methods applied are static and traditional. This study aimed to streamline these challenges encountered by clients and employees through the implementation a three-tier queue management system utilizing the agile model incorporating iterative phases of planning, design, development, and testing, with inputs from users' feedback. The evaluation comprises 115 participants composing supervisors, information and communication technology experts, staff, and service clients; mixed quantitative surveys with qualitative usability feedback for user-centric improvement using technology acceptance model and ISO-9126-4 frameworks. The results showed a consistently high score on key criteria, namely, perceived ease of use, structure & layout, relevance, functionality, interactivity and completeness with scores ranging from 4.65 to 4.80, which falls into the highly effective level. Findings represent the strength of the system and design strategy toward reduction of wait times, enhanced transaction flow and processes, and improved transparency. Further findings showed that the three-tier queuing system is beneficial to land transportation agency in fine-tunning their operational processes and increasing revenue.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Like many government agencies, Land Transportation hold the responsibility of providing quality services to people in various types [1] including issuance of licenses and permits that involves long queuing in the clients' arrival area. And, this has been a major issue faced by the Land Transportation Districts in Eastern Visayas, Philippines for several years due to prolong waiting time or queuing and the absence of transaction tracker display for the clients. These concerns cause customer dissatisfaction, overcrowding, resulting in backing out from the services, and revenue loss [2]. Such reasons inspire the researchers to conduct a study intending to resolve this major issues choosing LTO – San Juan District as the initial focus of the study.

The best queue should have shorter queues while serving the clients per unit of time because mismanaged queue becomes more challenging and causes dissatisfaction especially as data volume rises [3], [4]. Additionally, Robas et. al., found out that an effective queue management makes the customers feel better; thus, resulting in good service feedback [5]. Moreover, the LTO also uses non-flexible approach of monitoring transactions by giving paper tickets; and if misplaced risk of losing their spot is another concern. The queue system is used for having a systematic and faster process in transacting business because service clients who experience poor service delivery, especially when it entails waiting in long queues could result to departure without availing the services [3], [6].

And the speed of service, which is a major problem in the Philippine government, reflects how efficient the system is by which affects customer satisfaction [2]. The core aim of every institution is not to lose customers or clients by providing with the services that will satisfy their needs [3]. Adding unique features to a queuing application like automation of steps in the workflow through generation of automatic ticket identifiers [5], employee workstation, and client electronic display tracker boost transparency and accountability. Clients' experience in the queue such as ineffeciencies becomes a determining factor of their first impression of the organization and therefore can result to substantial losses [7]. Thus, reducing the unnecessary queuing time became a critical issue for the restaurant owners to promote customers' satisfaction [8]. Developing the three-tier queue management system has a main goal of linking the 3 systems namely client registration, employee workstation and client display.

This implemented system has the intention of reducing wait times, solving the dissatisfaction of service clients, streamlining employees' work performance and the use of resources, and providing a display tracker for the service clients to monitor the status of their requested services in the LTO – District Offices. Queueing system is influential to the supervision and the service quality which lead to revenue management strategies of an organization [9], [10]. Thus, this transformation addresses the increasing demand of the service clients particularly on transparency and wait time reduction. Moreover, queuing provides the cornerstone of efficiency as they assist employees and managers in tracking, prioritizing, and ensuring the delivery of services and transactions [7], [11]. Specifically, it sought to achieve the following objectives:

1. To assess the existing approach of servicing Land Transportation clients;
2. To propose possible enhancements in the existing strategy;
3. To solve the challenges encountered by the service clients and employees during license or permit applications from the service request to completion phases.

II. THEORETICAL REFERENCE

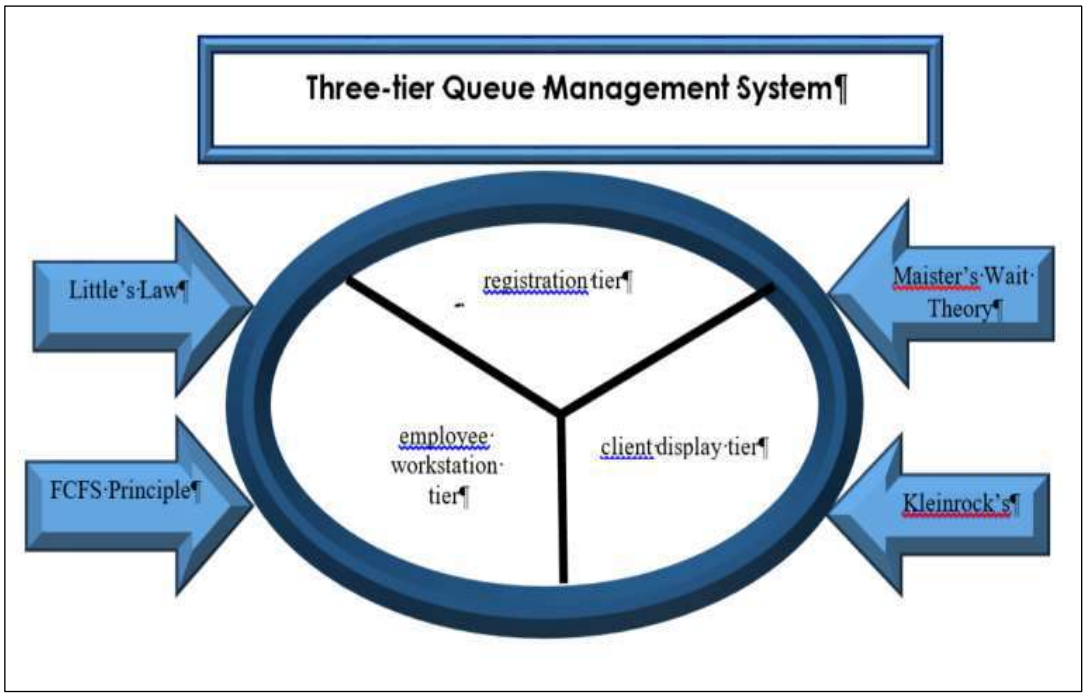


Figure 1: Theoretical Framework.
Source: Authors, (2026).

The three-tier queue management system was structured developed focusing the registration tier during the client's arrival for the selection of the service type and the issuance of the queue number an approach of minimizing bottleneck in the service area; employee workstation tier, and client display tier. In the case of the Land Transportation Districts of Eastern Visayas, Philippines, some of the customers or clients in the case of the Land Transportation leave without receiving the service maybe because they are tired of waiting the line [12] and status of their transaction cannot be tracked by the service clients.

Thus, Little's Law states that the average waiting time and the average number of items or clients (in the case of land transportation application) are important measurements in the decision making of the district supervisor [13]. First-Come, First-Serve Principle is also utilized in this system to ensure fairness by serving customers in the order they arrive or introduces and thereby promotes flexibility for exceptional situations or priority situations such as (Kleinrock, 1975) priorities to tasks or customers-based urgency or value. Moreover, the Service Quality and Perceived Wait Theory of Maister (1985) emphasize perceived waiting times and customer satisfaction are essentials for designing modern queuing systems [14].

The three tiers of the queue management system is implemented to manage service client's transaction's flow, promote quality service, and enhance client experience. Also, effectively controlled clients can avoid overcrowding and thus, the efficiency of the queue model may include important factors such as arrival time in the queue or server, average service time, average waiting time, and queue structure [8-10]. Transparency and real-time tracking of transactions through the display tier thereby enhancing customer satisfaction and ultimately improving overall operational efficiency.

III. MATERIALS AND METHODS

The development of the three-tier queue management system followed the agile model (see Figure 1) and occurred in two primary phases: development and evaluation. Development was iterative process that included: plan, design, develop, test, deploy, and review. This methodology was used because of its operational appropriateness, speed and the capability of adapting to the changing needs of the service clients and the personnel in the Land Transportation Offices (LTOs). In the planning stage, consultation with the key informants from Land Transportation District of San Juan, Southern Leyte, Philippines, District Supervisors, and Regional Administrators was held to elicit service clients' requirements. The consultation made it necessary to have a improved queue system for the Land transportation district offices to replace the traditional physical single-server waiting line to a three-tier queue management system to ensure efficient and fair government service delivery by reducing wait times in a queue lines, and thus, streamlining processes. This input and the fruits of extensive literature reviews shaped the goals for the system development including user-friendly graphical interfaces and a more structured approach to client service delivery. Prototyping in evolutionary refinements to increase scalability and usability was also conducted. Core functionalities include organization and management of clients' flow, minimizing wait times and improving the overall client journey across the different interactive touchpoints. These have been built and evolutionarily improved with the help of modern technology and agile development methodologies from stakeholders' feedback.

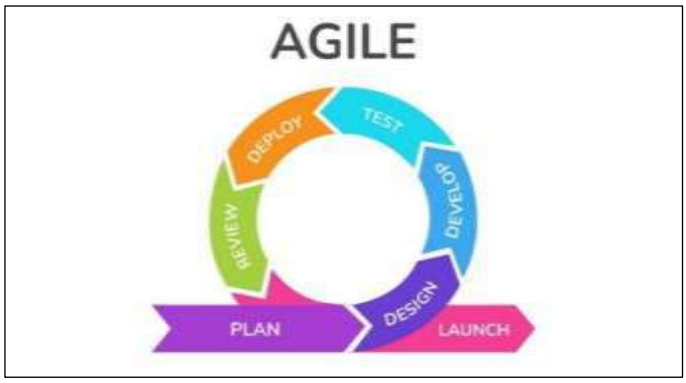


Figure 2: System development model.
Source: Authors, (2026).

On the other hand, the assessment phase used appropriate and detailed frameworks, including the TAM and ISO 9126-4 which laid a basis for assessing the system-based on a criterion that involved perceived ease of use, structure and layout, relevance, interactivity, functionality, and completeness. This phase involved 115 participants, including administrative officer, district supervisors, ICT experts, staff, and randomly selected service clients from LTO – San Juan district and other 19 district and satellite offices in Eastern Visayas, Philippines who completed a 20-item questionnaire adapted from Davis and Elling [15], [16]. According to Isaac et. al., [17], a sample size ranging from 100 to 200 is typically considered adequate for descriptive studies to characterize a population's attributes. To complement this quantitative data, usability testing, and feedback sessions offered qualitative insights into the system's performance. A month after the system's turnover, feedback was solicited to facilitate further enhancements and improvements.

IV. RESULTS AND DISCUSSIONS

A. System Development of the three-tier queue management system

Ease of doing business and efficient government service delivery are the bases of the researchers in identifying the activities involved in all the stages of system development and implementation. Identification of development models, management, hardware, software, graphical user interfaces, online forms and menus for integration were included. Streamlining processes, optimization of wait times, avoiding congestion, and efficiency in the delivery of government service are the intentions of this development. Hence, the queue management system provides simplified, centralized, and enhanced throughput derived analytically particularly to the land transportation service clients making it ready to address a range of real-world application [7].

This also allows the system administrator to monitor every stage, specifically in securing the queue number, status of the transactions in every service counter and up to the completion stage of permits or license application. The system was developed using various programming tools, including Visual Studio Code as source code editor, PHP and Rust programming languages for dynamic scripting, Rust with Tauri as framework for its frontend, Laragon for testing, CSS and Bootstrap for designing and lay outing of the elements and background, JavaScript for creating interactive and dynamic webpages, and MySQL for database, storing user/admin credentials and transaction details; TV/Monitor for the system's real-time display including queuing numbers and their corresponding counter locations; touch screen Monitor for transaction display and a thermal paper use for printing the queue numbers.

Figures 3 and 4 depicts the system architectures, explaining the technical setup of the three-tier queuing management system both for the service client and the employee assistant. These figures describe the entities and how they interact and communicate between the three-tier systems, employees and the service clients and all interface modules. The diagrams show smooth for diverse user groups such as system administrators, touch point employees, and clients because the system is scalable and robust. It also shows a straightforward working and also explains how the application supports real-time flow of processes, enhanced operational efficiency and interactivity.

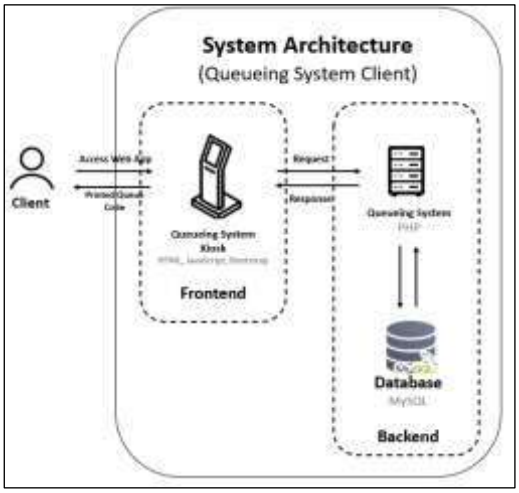


Figure 3: System Architecture for the Client.
Source: Source: Authors, (2026).

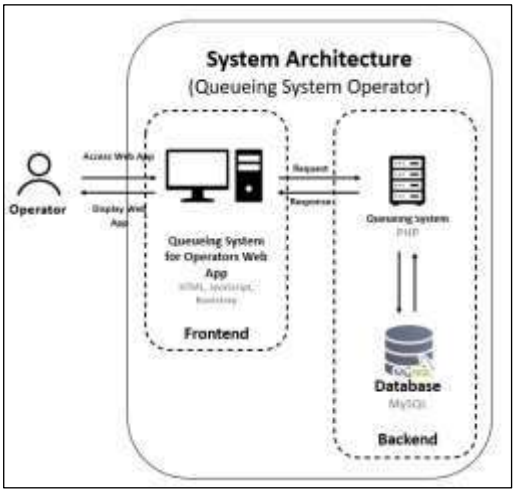


Figure 4: System Architecture for the Employee Assistant.
Source: Authors, (2026).

The interaction process between the service clients accessing the system and the step-by-step flow of the three-tier queue management system is demonstrated in Figure 5 where the service area has a digital kiosk that categorizes the services into Registration, Licensing, and LETAS. Fast printing speed and low maintenance are two major reasons of choosing a thermal printer connected to the kiosk for the generation of a queue number after the client's selection of the service type. Any previous data and information will be cleared for privacy reasons in the registration tier. Correspondingly, all the requested services with the queue numbers are dynamically posted and updated in the second and third tiers of the system called personnel workstation, display tier respectively. A series of rounds of tests were done to ensure that all the touch points were functional and usable.

These various testing operations were divided into three categories, namely: 1) unit testing –where hardware and software were integrated; 2) functional testing- where the LTO – San Juan personnel and service clients were allowed to operate for its functionality if it met their needs and expectations; 3) user acceptance testing- the clients and the employees operates the system to check its completeness and relevance to their needs and expectations; 4) bug fixing iteration where errors in the unit testing phase was noted and the system is revised according to these concerns; running the system is evolutionary until no errors were found. After a month, a review was conducted, where feedback from the personnel and the service clients was acknowledged and resolved.

Figure 6 shows the flow of the personnel tier system that started on the registration providing personal information such as User name-password combinations. The correctness of these information lead to user log in and displays the dashboard that contains number of active workstations, service categories, and personnel users. The transactions in the personnel station is updated in real-time and automatically updates the display tier system. Both audio and visual are the multi channels provided in this three-tier queue management system.

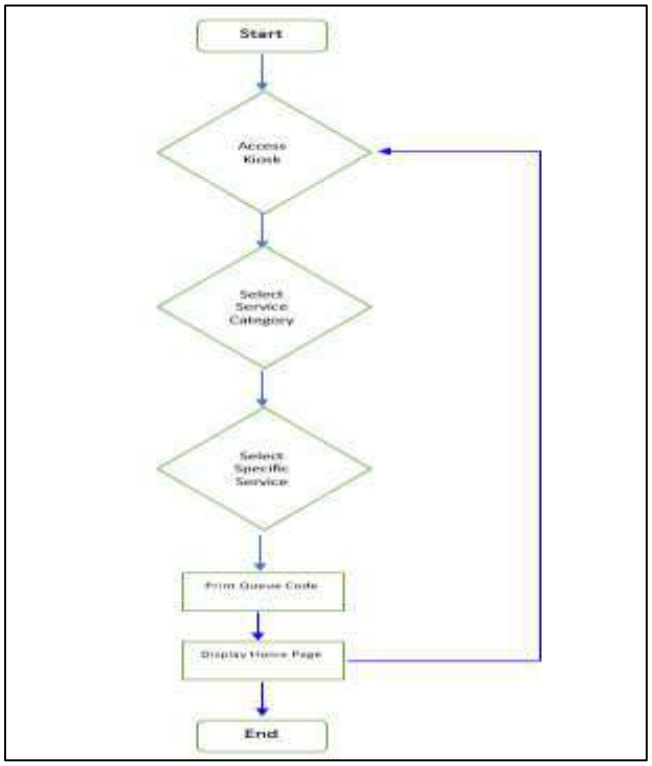


Figure 5: Flowchart Diagram for the Client.
Source: Authors, (2026).

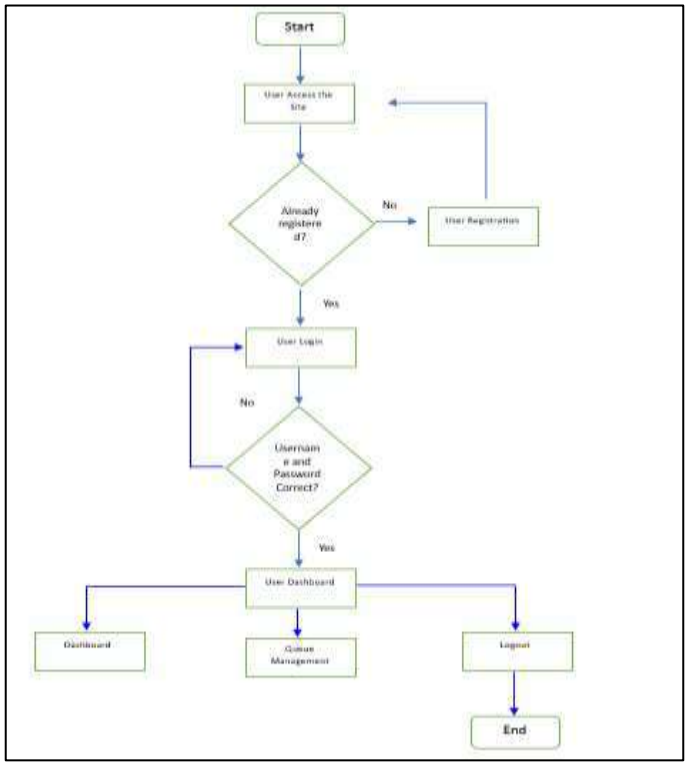


Figure 6: Flowchart Diagram for the Employee.
Source: Authors, (2026).

After the system was operational, it was piloted in San Juan Land Transportation District and in batches to other 19 districts and satellite offices for implementation. The researchers conducted feedback after a month and then scheduled a meeting with the district supervisors, IT experts, and other LTO personnel for enhancements. Hence, user acceptance is the top priority in the initial adoption stage. The final deployment was in full swing after fixing minor errors. This Three-tier Queuing System utilizes a Local Area Network (LAN) - based on web platforms using compatible browsers with interactive interfaces by inputting login credentials. It has limited human intervention, specifically in managing the queues while avoiding line jumpers and alleviating disputes within the queue area, increasing client satisfaction, and providing fast and quality services.

Also, clients feel empowered by being in control of their own time in an automated queuing system, which is directly connected to the quality of services [8], [18]. Waiting for service is a common phenomenon, and customer queuing time has been considered a key factor when evaluating a service-oriented institution, specifically the government [7]. Moreover, using the queuing system significantly affects seventy-eight (78) percent of the quality of service and supervision, considering the utilization rate and average waiting time [9]. Structure and layout of this system meets accessibility standards particularly on the screen reading compatibility of the service clients. Figure 7 displays all the LTO services such as Licensing, Registration and LETAS that are configured and controlled by the admin selected from the personnel appropriately and information technology expert.

The breakdown of these services is handled by the three-tier queue management system including issuance of drivers and conductor licenses, registration and renewal of vehicle, and the issuance of related documents (official receipt, certificate of registration, and other traffic violations). Congestion prevention is one of the most important factors that is considered in the development of this system. Figure 8 is a sample of a unique identification number printed from a thermal printer during the arrival of the waiting area which details the service category that is requested by the client. Also, increasing the number of servers causes a decrease in the average waiting time, thereby increasing service clients satisfaction [19]. However, queue path can be expanded from 1 to 2 with the use of the queue number so that the length of the queue decreases its service time in each service counter [20].



Figure 7: Client transact.
Source: Authors, (2026).

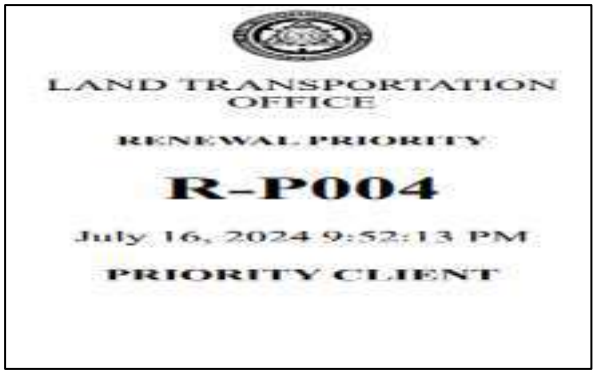


Figure 8: Client Queuing Number.
Source: Authors, (2026).

All operational transactions, namely, number of workstations, the number of active service categories, and the number of personnel users, are displayed in Figure 8 which is the Main Interface of the Queuing System, where each district logs in independently. Each district can navigate the system independently, including Licensing, Registration, and Letas services. In addition, waiting time was minimized from 1 hour per client to 20 minutes. Client satisfaction was noted due to least wait times while maximizing revenue and increasing the productivity among employees [4]. It also enables LTO– -San Juan personnel to efficiently manage clients' flow by getting the queue number and waiting time, monitoring the clients' applications, and controlling the influx of new and existing clients for vehicle registration, paying penalties, and applying for licenses.

Figure 9 is the log in page, where considered entry point of all admin or employee users. It is user-friendly and has a secure authentication method, such as the username-password combination. This serves as the interface where admin or employee users begin interacting with the three-tier queue management system displaying the dashboard offering personalized access to service clients data while ensuring security and management. The functionality of this system prevents unauthorized access as designated by the role-based of the system administrator.



Figure 9: LTO Personnel.
Source: Authors, (2026).



Figure 10: Home.
Source: Authors, (2026).



Figure 11: Transaction.
Source: Authors, (2026).

Figure 10 is the Home page that displays the number of workstations, service categories, and employee users that are updated in real-time. Transaction details, including queue number and station assignment per employee user, are managed and controlled by the system administrator. Each workstation or counter contains a click-enabled button to represent a transaction with corresponding queue number per client that are served, called, forwarded, finished, and discarded. If one counter becomes overloaded, queue numbers will be reassigned to other available counters.

Figure 11 is a sample of the transaction page in the three-tier queue management system containing a real-time and visually enticing interface with which employee users could engage queue controls such as call service clients, forward, finish or complete transactions, discard/cancel and hold or pause service done in multichannel approach such as audio and digital. Queue entries in a specific workstation including complete transaction history can also be accessed to efficiently optimize client services and provides next steps of every transaction for easy re assignment of queue numbers and counters. This design embodies current principles of modern queuing system that is simple and well-organized that are in compliance with data protection regulations.

To ensure an efficient and transparent servicing of the clients, a call interface in the employee workstation was implemented as depicted in Figure 12. This interface notifies the service clients in their turn for service and applies dynamic routing to the correct workstation or counter through audio and display monitor announcements. To optimized transparency and unnecessary inquiries, all active workstations are displayed in the third-tier panel as shown in Figure 13. This display updates transactions in real-time that are called, forwarded, or completed. Also, this panel prevents from the occurrence of a dispute by providing a digital summary visible to service clients**.**



Figure 12: Workstation Call Interface.
Source: Authors, (2026).



Figure 13: Client Display Tier.
Source: Authors, (2026).

B. Evaluation of the Three-tier Management System

In the evaluation phase of the system as shown in Table 1, LTO – administrative officer, district supervisors, ICT experts, staff and the service clients consistently with high ratings across all variables. The mean scores for perceived ease of use, hyperlink, structure and layout, relevance, functionality, interactivity, relevance, and completeness were well at the very high effectiveness level between 4.65 and 4.80. These ratings show the system's performance is very good since it meets the users' expectations in different criteria. Finally, this indicate that the respondents had a high level of concurrence with one another towards the overall effectiveness of this system.

Table 1: Summative evaluation of the three-tier queue management system across different respondent groups.

Variables	Mean	Min	Max	Effectiveness Level
Perceived ease of use	4.65	3.00	5.00	Very high
Structure and Layout	4.65	3.00	5.00	Very high
Relevance	4.80	3.00	5.00	Very high
Interactivity	4.70	3.00	5.00	Very high
Functionality	4.70	3.00	5.00	Very high
Completeness	4.70	3.00	5.00	Very High
Note: 1.00-1.79 Very low; 1.80-2.59 – Low;2.60-3.39 – Average; 3.40-4.19 – Moderately high; 4.20-5.00 – Very high				

Source: Authors, (2026).

All the variables are uniform with minimum and maximum scores ranging from 4.65 to 4.80. This range suggests that the system can provide a good experience to most users while flagging areas for improvement. Relevance is the highest rated with M=4.80, implying that the system can stay in line with what its users need and eventually expect. Similarly, consistency of high rankings in completeness, functionality and interactivity indicates that its necessary modules are practically functioning and fosters dynamic exchange between the user and the systems. Strong scores across all dimensions mean that there has been a robust design and usability of the system in the land transportation's operational transactions. However, perceived ease of use, structure and layout with at least 4.65, mean it should strive at further refinements what could answer some of the evaluation from the users and ensure that the system remains helpful and responsive to diverse clients' need. The findings also suggest system's strengths and offer insights into what needs to be improved to ensure that diverse users have expectations met so that land transportation services it provides are enhanced through innovative digital solutions.

Moreover, the findings also revealed that the customer waiting time is reduced from an hour to 20 minutes and the negative impression was alleviated thereby increases client's satisfaction, fairness and transparency. The development and implementation of this three-tier queue management system is aligned with the mandate and principles of LTO towards fair and transparent services to clients. It also fosters trust and confidence in LTO personnel by reducing clients' conflict and faster services. Lastly, the queuing system enhances the client's ease of accessing the system and solves the challenges encountered by the current traditional queuing approach. More importantly, decreasing the average waiting time of service clients in a queue, increases satisfaction [19], [20]. To have a uniform queuing system for the LTO - Eastern Visayas Region, Philippines, installation and implementation was done to other 19 districts and satellite offices with the approval of the regional director since these offices encountered identical challenges with LTO – San Juan District.

V. CONCLUSIONS

The developed queuing system is very significant to LTO supervisors, personnel and the service clients of the region because it enables them to improve users' satisfaction and enhanced operational efficiency. The reduction of wait wait times from the registration to the completion of the transaction increases employee productivity resulting to more satisfied clients leading to increased revenue. This system provides positive experiences to service clients and word of mouth referrals, thereby stronger agency reputation. Furthermore, the functionality aspect of the system aids the supervisors in having real-time monitoring and reporting. Interactivity of the system interfaces creates an organized communication between the users and the system making all the services fast and functional. Finally, management of the flow of clients' transactions are straightforward, facilitating faster personnel assistance, appropriate allocation of available resources, and streamlining the entire land transportation operations. Finally, this queuing system was developed in a three-tier design for future expansion.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Ines G. Falcon, Kenneth Jay L. Dugaria and Danna C. Remojo.
Methodology: Ines G. Falcon, Kenneth Jay L. Dugaria.
Investigation: Ines G. Falcon, Kenneth Jay L. Dugaria
Discussion of results: Ines G. Falcon, Kenneth Jay L. Dugaria and Danna C. Remojo
Writing – Original Draft: Ines G. Falcon
Writing – Review and Editing: Ines G. Falcon and Kenneth Jay L. Dugaria.
Resources: Kenneth Jay L. Dugaria
Supervision: Kenneth Jay L. Dugaria and Danna C. Remojo.
Approval of the final text: Ines G. Falcon, Kenneth Jay L. Dugaria and Danna C. Remojo.

VII. ACKNOWLEDGMENTS

The authors would like to thank the participation of the personnel and the randomly selected service clients. The authors also would like to express their recognition to the head of research, innovation, and extension services of one of the university's satellite campuses for the guidance and constructive analysis of the result of the study.

VIII. REFERENCES

[1] Solivio, M. P., Bare, M. A. D., & Castro, F. M. F. An Adaptation of Industry 4.0: Modifying the Implemented Queuing System in Philippine Statistics Authority Applying Monte Carlo Simulation. Proceedings of the International Conference on Industrial Engineering and Operations Management Monterrey, Mexico, November 3-5, 2021

[2] Gimba, U. A., Okoronkwo, C. D., Yusuf, M., Musa, A. S., & Ali, M. S. (2020). Queue monitoring system for bank. *Dutse Journal of Pure and Applied Sciences (DUJOPAS)*, *6*(2), 269-276.

[3] Dudin, S. A., Dudina, O. S., & Kostyukova, O. I. (2023). Analysis of a Queuing System with Possibility of Waiting Customers Jockeying between Two Groups of Servers. *Mathematics*, *11*(6), 1475.

[4] Moosavi, S. A. H., Kuisma, M., Saeed, M. O., Appiah, E., Jama, M., Iqbal, Z., ... & Kaura, D. (2018, March). Saffara: Intelligent queuing application for improving clinical workflow. In *Qatar Foundation Annual Research Conference Proceedings Volume 2018 Issue 3* (Vol. 2018, No. 3, p. ICTPD851). Hamad bin Khalifa University Press (HBKU Press).

[5] Robas, K. P. E., Solivio, M. P., Haboy, D. A. A., Taw, C. A. Y., & Borres, R. D. (2021). An Adaptation of Industry 4.0: Comparison between the Implemented Queuing System of Land Transportation Office and Proposed System Applying Monte Carlo Simulation, Lean Management, and Analytical Hierarchy Process Method. In *the International Conference on Industrial Engineering and Operations Management*.

[6] Balasa, J. A., Jardiolin, R. H. , Villareal, J. M. , Dequilla, R. J. & Tady, R. , Jr. (2018). College of Computer Studies queuing and electronic bulletin system (Unpublished special paper). Central Philippine University, Jaro, Iloilo City.

[7] Alnowibet, K. A., Khireldin, A., Abdelawwad, M., & Mohamed, A. W. (2022). Airport terminal building capacity evaluation using queuing system. *Alexandria Engineering Journal*, *61*(12), 10109-10118.

[8] Ghazali, N. A., & Amit, N. (2023, January). Comparing the Efficiency of Two Queuing Models for a Fast-Food Restaurant Using Analytical Queuing Theory In Fundamental and Applied Sciences in Asia: International Conference on Science Technology and Social Sciences (ICSTSS 2018) (pp. 179-186). Singapore: Springer Nature Singapore.

[9] Kim, K., Kim, M. J., & Jun, J. K. (2020). Small queuing restaurant sustainable revenue management. *Sustainability*, *12*(8), 3477.

[10] Sudarman, D. (2023). THE INFLUENCE OF QUEUEING SYSTEM AND SUPERVISION ON SERVICE QUALITY IN JAKARTA MRT TRANSPORTATION. *CASHFLOW: CURRENT ADVANCED RESEARCH ON SHARIA FINANCE AND ECONOMIC WORLDWIDE*, *2*(3), 461-473.

[11] Cowdrey, K. W., de Lange, J., Malekian, R., Wanneburg, J., & Jose, A. C. (2018). Applying queueing theory for the optimization of a banking model. *Journal of Internet Technology*, *19*(2), 381-389.

[12] Shortle, J. F., Thompson, J. M., Gross, D., & Harris, C. M. (2018). *Fundamentals of queueing theory*. John Wiley & Sons.

[13] Little, J. D., & Graves, S. C. (2008). Little's law. In *Building intuition: insights from basic operations management models and principles* (pp. 81-100). Boston, MA: Springer US.

[14] Bassey, J. U., Boniface, O. P., Abiji, A. E., & Chidubem, B. S. (2025). Effects of waiting line management on customer satisfaction: A Study of selected banks in Ogoja, Nigeria.

[15] Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly. 1989;13(3):319-40. https://doi.org/10.2307/249008.

[16] Elling S, Lentz L, De Jong M. Website evaluation questionnaire: development of a research-based tool for evaluating informational websites. In Electronic Government: 6th International Conference, EGOV 2007, Regensburg, Germany, September 3-7, 2007. Proceedings 6 2007 (pp. 293-304). Springer Berlin Heidelberg.

[17] Isaac S, Michael WB. Handbook in research and evaluation: A collection of principles, methods, and strategies useful in the planning, design, and evaluation of studies in education and the behavioral sciences. 3rd ed. San Diego, CA: EdITS Publishers; 1995.

[18] Wu, C. H., Yang, D. Y., & Lin, W. L. (2025). Modeling and cost optimization of controlling customer arrivals in queueing systems with two waiting areas. *Journal of Industrial and Management Optimization*, *21*(2), 1251-1268.

[19] Ibeawuchi, N. U., & Tsetimi, J. (2023). ANALYSIS OF THE TICKETS TERMINAL OF KADUNA RAILWAY STATION: A QUEUING THEORY APPROACH. *Nigerian Journal of Science and Environment*, *21*(1).

[20] Raicu, S., Costescu, D., & Popa, M. (2023). Effects of the Queue Discipline on System Performance. *AppliedMath*, *3*(1), 37-48.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 567-573. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2790

RESEARCH ARTICLE OPEN ACCESS

STRUCTURAL OPTIMIZATION USING GREY WOLF OPTIMIZER AND ITS VARIANTS

Bhavik D. Upadhyay*[1], Sonal T. Dave[2], Vishal J. Pandya[3], Jignesh G. Parmar[4] Hasmukh P. Koringa[5], Miral J. Patel[6]

[1]Associate Professor, Department of Mechanical Engineering, Shantilal Shah Engineering College, Bhavnagar, Gujarat, India,
[2]Lecturer, Department of Mechanical Engineering, Sir Bhavsinhji Polytechnic Institute, Bhavnagar, Gujarat, India.
[3,4]Assistant Professor, Department of Mechanical Engineering, Shantilal Shah Engineering College, Bhavnagar, Gujarat, India.
[5,6]Associate Professor, Department of Electronics & Communication Engineering, Government Engineering College, Bhavnagar, Gujarat, India,

[1]http://orcid.org/0000-0001-7490-8925, [2]http://orcid.org/0009-0001-1910-7188, [3]http://orcid.org/0000-0002-3467-4826, [4]http://orcid.org/0000-0002-8533-8304, [5]https://orcid.org/0000-0002-2521-3972, [6]https://orcid.org/0000-0003-0047-9552

Email: *bhavik1483@gmail.com, davesonal83@gmail.com, visheepandya@gmail.com, jigar.mechanical1@gmail.com, hpkoringa@gmail.com, mjpatelgec@gmail.com

ARTICLE INFO

Article History
Received: October 8, 2025
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
structural optimization,
Meta-heuristics,
GWO,
mGWO,
vwGWO.

ABSTRACT

Now a days, many metaheuristic optimization algorithms are invariably used in the field of engineering and data science. In structural mechanics, many stochastic optimization algorithms are employed for solving the complex numerical problems. Suitable optimization algorithms must be selected as the structural mechanics problems are often having non-linear objective function, constraints are having complex behavior and many design variables are involved in it. In present study, the capability and competency of standard grey wolf optimizer (GWO) with modified grey wolf optimizer (mGWO) and grey wolf optimizer with variable weight (vwGWO) is observed through several benchmarked numerical case studies. To control the geometric constraints, penalty method is employed in the study. The performance of the standard GWO algorithm is compared with the results obtained by the modified grey wolf optimizer and grey wolf optimizer with variable weight.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Since last three decades, metaheuristic optimization techniques are evolved very fast with prominent features. Some of them are very popular techniques like, Genetic Algorithm (GA) [1], Particle Swarm Optimization (PSO) [2], Artificial Bee Colony (ABC) [3], Ant Colony Optimization (ACO) [4], Grey Wolf Optimizer (GWO) [5] etc., not only in the field in computer science but also in the field of engineering, technology and structure mechanics as well. The metaheuristic optimization algorithms are very simple to implement and also flexible.

They are mostly inspired by simple concepts and this allows the researchers to simulate the different concepts from nature, to propose newer metaheuristic approach, to hybridize two or more metaheuristic algorithms or to improve the existing metaheuristic optimization algorithms [5]. Many researches were also performed in the field of optimization considering various case studies and their results are compared [6]. In article, Grey wolf optimizer explained briefly in section II followed by its variants, mGWO and vwGWO in sections III and IV respectively. Three case studies are discussed implementing GWO, mGWO and vwGWO in section V. The competency of the GWO, metaheuristic algorithm with mGWO and vwGWO is discussed in section VI.

II. GREY WOLF OPTIMIZER

Grey wolf optimizer is proposed by [5] which is based on the social and hunting behavioral of grey wolves. Grey wolves are apex predators belongs to Canidae family and hence they are at the top of the food chain. Grey wolves are living in groups. The group size on an average of the wolves is generally 5 to 12.

Grey wolf optimizer imitates the social behavior and hunting mechanism of the grey wolves. The optimum value search mechanism of the algorithm is based on the three steps of hunting technique adopted by the grey wolves:

- Tracking, chasing and approaching the prey.
- Pursuing, encircling and harassing the prey until it stops the motion.
- Attacking the prey.

The mathematical model of the grey wolf optimizer is outlined as follows:

II.1 SOCIAL HIERARCHY

The optimum solution or the fittest solution of the problem is decided as per the social hierarchy of the wolves in a pack to model the optimization algorithm mathematically. In a pack of wolves, the leader wolf is considered as α wolf. The position of the α wolf is noted as fittest solution from the possible solutions. The second-best solution is considered as the position of β wolves which are the subordinate wolves to leader wolf and help in decision making and other group activities. The lowest ranking of the wolves in a pack is known as ω. These are the wolves who play the role of scapegoat or sometimes work as baby sitters. The third best solution is considered as the position of δ wolves. If the wolf is not α wolf, β wolf or ω wolf, he/she is considered as δ wolf. This type of wolves scouts the pack and looking the boundary of the territory. In case of any danger, they warn the pack.

II.2 ENCIRCLING THE PREY

During hunting, the wolves encircle the prey. The encircling behavior of the wolves is modelled mathematically as follows:

$$\vec{D} = \left|\vec{C}.\overrightarrow{X_p}(i) - \vec{X}(i)\right| \quad (1)$$

$$\vec{X}(i+1) = \overrightarrow{X_p}(i) - \vec{A}.\vec{D} \quad (2)$$

$$\vec{X}(t) = \frac{\overrightarrow{X_1} + \overrightarrow{X_2} + \overrightarrow{X_3}}{3} \quad (3)$$

Where, i indicates the current iteration. $\vec{A}$ and $\vec{C}$ are the vectors coefficients. $\overrightarrow{X_p}$ indicates the position of the prey and $\vec{X}$ is the position vector of the grey wolf. The vectors coefficients $\vec{A}$ and $\vec{C}$ are calculated as per following expressions:

$$\vec{A} = 2\vec{a}\,.\vec{r}_1 - \vec{a} \quad (4)$$

$$\vec{C} = 2\,.\vec{r}_2 \quad (5)$$

Here, $\vec{r}_1$ and $\vec{r}_2$ are arbitrary vectors in the domain of [0,1] and $\vec{a}$ is the decrement vector which varies linearly from 2 to 0.]

II.3 HUNTING TECHNIQUE

In hunting of the prey, the hunt is generally guided by the α wolf. β wolves and δ wolves also participate in hunting. Initially, there is no idea about the optimum(prey) position in a search space and hence to model the mechanism mathematically, the initialization of the hunt by the wolves in the search space are recorded as initial solution and the positions assigned as per the fitness of the solution. The best initial solution is recorded as α-position, second best solution is recorded as β-position and third best solution is recorded as δ-position.

General steps of GWO algorithm as follows:

Grey wolf pack formation
Initialize the position of wolves
Fitness calculation
Position record
Repeat
Update the random parameters
Fitness calculation
Position record appraisal
α-wolves (the best solution),
β-wolves, (the second best solution)
δ-wolves (the third best solution)

until
Convergence criterion achieved

The controlling parameter ‘a’ reduces linearly from 2 to 0 when the iterations are being carried out.

$$a = 2\left(1 - \frac{i}{N}\right) \tag{6}$$

Where, i is the $iteration$ and N is $maximum\ number\ of\ iterations$.

III. MODIFIED GREY WOLF OPTIMIZER (mGWO)

Modified grey wolf algorithm is proposed by [7] mainly focused on the good balance between proper exploration and exploitation to find the global optima of real problems. In standard GWO algorithm, the coefficient vector $\vec{A}$ randomly varies between '-a' to 'a' and simultaneously over the course of iterations, $\vec{a}$ varies linearly from 2 to 0. Balancing between $\vec{A}$ and $\vec{a}$ needs much attention and fine adjustment is required to avoid the solution falls under local optima. To overcome this issue, the researchers suggested to use exponential decay function for tuning the parameter 'a'. Earlier, In standard GWO, the value of 'a' decreases from 2 to 0 linearly as per equation (6). as per the suggested exponential decay function (Mittal N. et al 2016) the expression is as follows:

$$a = 2\left(1 - \frac{i^2}{N^2}\right) \tag{7}$$

Where, i is the $iteration$ and N is $maximum\ number\ of\ iterations$. The suggested approach is adopted in the proposed research and applied to the benchmarked problems in section V to derive the competency and worthiness of mGWO.

IV. GREY WOLF OPTIMIZER WITH VARIABLE WEIGHT (vwGWO)

To make the GWO algorithm very efficient to handle the variety of problems with multi variables and high complexity of the real-life problems of engineering, a new hypothesis related to search mechanism of the grey wolf optimizer by Gao Z. M. et al in 2019 [8] is proposed as improved grey wolf optimization with variable weight (vwGWO). In standard GWO algorithm as per equation (2.3), dominants play an equal role in search mechanism of grey wolf optimizer. During exploration process, in the initial phase of algorithm, the alpha wolf (α) has the better position than β wolf or δ wolf in each iteration cycle. In searching process α wolf is the better contributor so his/her position should be given weight more than the positions of β wolf and δ wolf in next search cycle [8].

$$\vec{X}(t+1) = w_1\overrightarrow{X_1} + w_2\overrightarrow{X_2} + w_3\overrightarrow{X_3} \tag{8}$$

$$w_1 + w_2 + w_3 = 1 \tag{9}$$

The weights w_1, w_2 and w_3 are associated with α wolf, β wolf and δ wolf respectively and must satisfy the condition, $w_1 \geq w_2 \geq w_3$. For vwGWO, the proposed equations for variable weights are as follows:

$$\begin{aligned} w_1 &= \cos\theta, \\ w_2 &= \frac{1}{2}\sin\theta\cos\varphi, \\ w_3 &= 1 - w_1 - w_2 \end{aligned} \tag{10}$$

Where,

$$\theta = \frac{2}{\pi}\arccos\frac{1}{3}\cdot\arctan(i)\,; \tag{11}$$

$$\varphi = \frac{1}{2}\arctan(i) \tag{12}$$

V. NUMERICAL CASE STUDIES

Three cases are discussed to establish the competence of the GWO with its variants i.e. mGWO and vwGWO. MATLAB code is prepared for each numerical case and 10 runs are taken for varying search agents and number of iterations.

Case: I CANTILEVER BEAM

Problem statement:

Figure 1 shows stepped hollow cantilever beam with constant thickness, of square cross section. It is subjected to a vertical load (P) at free end of the beam structure. The problem is of weight minimization and design variables are height of different beam sections. The thickness of the beam elements is fixed as t = 2/3.

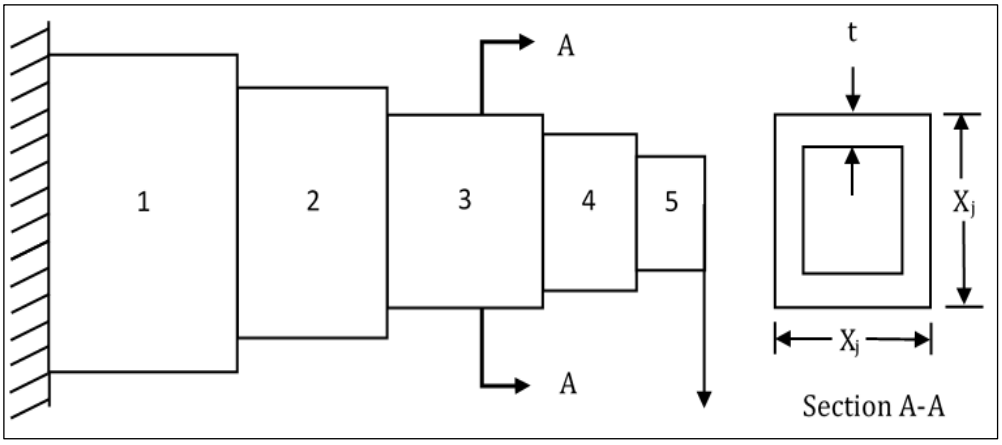


Figure 1: Cantilever Beam.
Source: Authors, (2026).

***Objective Function*:** Minimization

$$f(x) = f(x_1, x_2, x_3, x_4, x_5) = 0.0624\,(x_1 + x_2 + x_3 + x_4 + x_5)$$

***Constraints*:**

$$C = \frac{61}{{x_1}^3} + \frac{37}{{x_2}^3} + \frac{19}{{x_3}^3} + \frac{7}{{x_4}^3} + \frac{1}{{x_5}^3} - 1 \leq 0$$

***Range*:**

$$0.01 \leq x_j \leq 100$$

Table 1: Comparison of the results for cantilever beam problem.

	GWO	**mGWO**	**vwGWO**
X_1 (Element – 1)	6.0465	5.9381	6.0199
X_2 (Element – 2)	5.2891	5.3576	5.2858
X_3 (Element – 3)	4.4449	4.5531	4.5513
X_4 (Element – 4)	3.5429	3.4594	3.5022
X_5 (Element – 5)	2.1544	2.1772	2.1177
$F_{min.}$	1.3402	1.3407	1.3402

Source: Authors, (2026).

It is noticed that to optimize the weight of stepped cantilever beam, GWO requires 25 wolves and 50 iterations to reach the optimum function value. mGWO requires 25 wolves and 100 iterations to reach the optima and vwGWO needs 25 wolves and 100 iterations to reach the optimum solution.

Case: II TUBULAR COLUMN DESIGN

Problem statement:

Figure 2 shows tubular column section which carries a vertical load P. The column is joined by means of hinged joints at both the ends. The mean dia. of the column is kept between 2 cm to 14 cm and the thickness ranged between 0.2 cm to 0.8 cm. The induced stress in the tubular column should not be more than the buckling stress and the yield stress. The total cost of the column considering both, material cost and construction cost, can be taken as 5W + 2d, where W stands for weight in kgf and d is mean dia. of the column in cm. The cost of the component is taken as objective of the problem which should be minimize. The material property and related particulars of the column is given below:

P = 2500 kgf (compressive);
l = 250 cm;
Yield stress = 500 kgf/cm^2;
Modulus of Elasticity= 0.85 X 106 kgf/cm^2;
Weight density = 0.0025 kgf/cm^3;

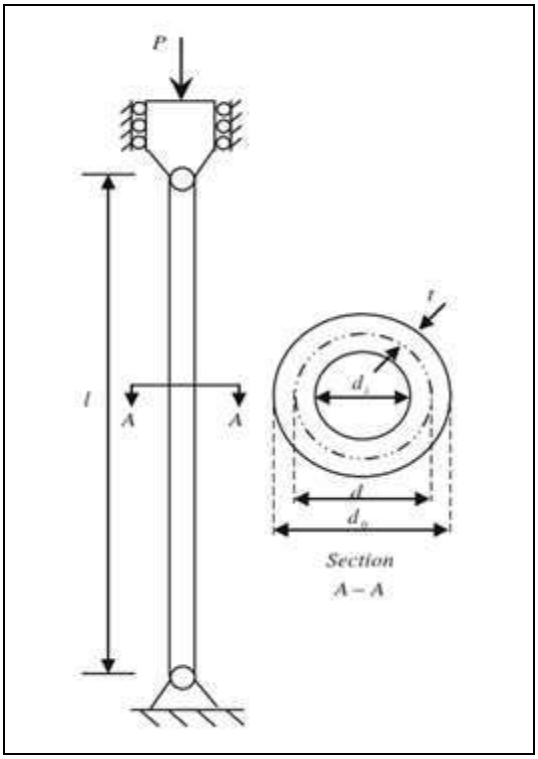


Figure 2: Tubular column design.
Source: Authors, (2026).

***Objective Function*:** Minimization

$$f(x) = f(x_1, x_2) = 9.82\, x_1 x_2 + 2x_1$$

Where,

$$X = \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} d \\ t \end{Bmatrix}$$

***Constraints*:**

$$c_1 = \frac{2500}{\pi x_1 x_2} - 500 \leq 0$$

$$c_2 = \frac{2500}{\pi x_1 x_2} - \frac{\pi^2 (0.85 \times 10^6)({x_1}^2 + {x_2}^2)}{8(250)^2} \leq 0$$

$$c_3 = \frac{2}{x_1} - 1 \leq 0$$

$$c_4 = \frac{x_1}{14} - 1 \leq 0$$

$$c_5 = \frac{0.2}{x_2} - 1 \leq 0$$

$$c_6 = \frac{x_2}{0.8} - 1 \leq 0$$

***Range*:**

$$2 \leq x_1 \leq 14$$

$$0.2 \leq x_2 \leq 0.8$$

Table 2: Comparison of the results for tubular column problem.

	GWO	**mGWO**	**vwGWO**
d (mean diameter)	5.4511	5.4501	5.4495
t (thickness)	0.29204	0.29222	0.29229
$F_{min.}$	26.5351	26.54	26.5404

Source: Authors, (2026).

It is noted that to optimize tubular column design, standard GWO requires 25 wolves and 200 iterations to reach the optimum function value. mGWO requires 25 wolves and 100 iterations to reach the optima and vwGWO needs 25 wolves and 200 iterations to reach the optimum solution.

Case: III THREE BAR TRUSS

Problem statement:

Figure 3 presents three bar truss. The objective is to find the optimum cross sectional area, to minimize the volume of the three bar truss structure. Each member length of the structure is 100 cm. The structure is subjected to load (F) of 2 kN/cm^2 and the permissible stress (σ) is not exceed to 2 kN/cm^2.

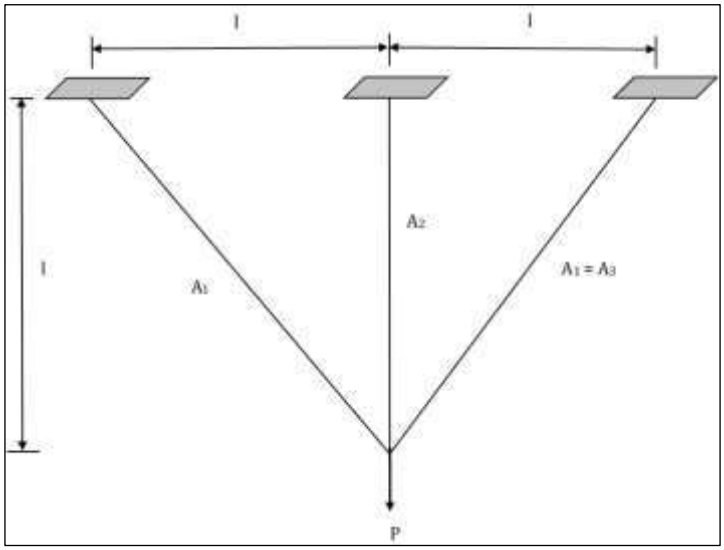


Figure 3: Three bar truss.
Source: Authors, (2026).

***Objective Function*:** Minimization

$$f(x) = f(x_1, x_2) = \left(2\sqrt{2}x_1 + x_2\right) \times l$$

Where,

$$X = \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} A_1 \\ A_2 \end{Bmatrix}$$

***Constraints*:**

$$c_1 = \frac{\sqrt{2}x_1 + x_2}{\sqrt{2}{x_1}^2 + 2x_1x_2} F - \sigma \leq 0$$

$$c_2 = \frac{x_2}{\sqrt{2}{x_1}^2 + 2x_1x_2} F - \sigma \leq 0$$

$$c_3 = \frac{1}{x_1 + \sqrt{2}x_2} F - \sigma \leq 0$$

***Range*:**

$$0 \leq x_1, x_2 \leq 1$$

Table 3: Comparison of the results for three bar truss structural problem.

	GWO	**mGWO**	**vwGWO**
A_1 (C/S area of member 1 & 3)	0.78776	0.78453	0.78947
A_2 (C/S area of member 2)	0.41085	0.42026	0.40601
$F_{min.}$	263.897	263.925	263.8969

Source: Authors, (2026).

It is noted that to optimize the cross-sectional area of the members of three bar truss structure, GWO requires 25 wolves and 100 iterations to reach the optimum function value. mGWO requires 20 wolves and 100 iterations to reach the optima and vwGWO needs 25 wolves and 100 iterations to reach the optimum solution.

VI. CONCLUSION

The present work explores the capability of metaheuristic optimization algorithms, GWO and also employed its modified versions known as modified GWO (mGWO) and improved GWO with variable weight (vwGWO). In case of stepped cantilever beam, GWO is found competent enough against mGWO and vwGWO whereas for tubular column design and three bar truss, mGWO is found competent in terms of less search agents and minimum iterations required to converge the solution. However, the effect of the upgraded versions i.e. mGWO and vwGWO is effective in case of multi-objective multiple constraints wherein the various variables affect the domain differently in the domain simultaneously.

VII. FUNDING

There are no financial implications from any sources or funding agency.

VIII. AUTHOR CONTRIBUTIONS

Conceptualization: Bhavik D. Upadhyay, Sonal T. Dave, Vishal J. Pandya, Jignesh G. Parmar, Hasmukh P. Koringa and Miral J. Patel.
Methodology: Bhavik D. Upadhyay and Sonal T. Dave.
Investigation: Bhavik D. Upadhyay and Sonal T. Dave.
Data collection and Analysis: Bhavik D. Upadhyay and Sonal T. Dave.
Discussion of results: Bhavik D. Upadhyay, Miral J. Patel and Sonal T. Dave.
Manuscript Writing – Original Draft: Bhavik D. Upadhyay and Sonal T. Dave.
Writing – Review and Editing: Bhavik D. Upadhyay, Hasmukh P. Koringa and Vishal J. Pandya.
Overall formatting: Vishal J. Pandya, Hasmukh P. Koringa and Jignesh G. Parmar.
Resources: Vishal J. Pandya, Miral J. Patel and Jignesh G. Parmar.

IX. REFERENCES

[1]Bonabeau E, Dorigo M, Theraulaz G. Swarm intelligence: from natural to artificial systems: OUP USA; 1999.

[2]Kennedy J, Eberhart R. Particle swarm optimization, in Neural Networks, 1995. In: Proceedings, IEEE international conference on; 1995. p. 1942–1948.

[3]Basturk B, Karaboga D. An artificial bee colony (ABC) algorithm for numeric function optimization. In: IEEE swarm intelligence symposium; 2006. p. 12–4.

[4]Dorigo M, Birattari M, Stutzle T. Ant colony optimization. Comput Intell Magaz, IEEE 2006;1:28–39.

[5]Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. Advances in engineering software, 69, 46-61.

[6]Upadhyay, Bhavik D., Sunil S. Sonigra, and Sachin D. Daxini. "Structural design optimization using grey wolf optimizer." Challenges and Opportunities in Industrial and Mechanical Engineering: A Progressive Research Outlook. CRC Press, 2024. 429-441.

[7]Mittal, N., Singh, U., & Sohi, B. S. (2016). Modified grey wolf optimizer for global engineering optimization. Applied Computational Intelligence and Soft Computing, 2016.

[8]Gao, Z. M., & Zhao, J. (2019). An improved grey wolf optimization algorithm with variable weights. Computational Intelligence and Neuroscience, 2019.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 574-580. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2829

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

AREA AND POWER EFFICIENT SPARSE AWARE ARCHITECTURE FOR POINTWISE CONVOLUTION UNIT USING VARIABLE DENSITY BOUND BLOCK

Golajapu Jyothsna[1], Narayanam Balaji[2]

[1, 2]Department of Electronics and Communications Engineering, Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India

[1]https://orcid.org/0009-0007-3515-907X, [2]https://orcid.org/0009-0003-5885-6688

Email: golajapujoe@gmail.com, narayanambalaji@jntucek.ac.in

ARTICLE INFO

Article History

Received: October 17, 2025

Revised: December 20, 2025

Accepted: January 15, 2026

Published: February 28, 2026

Keywords:

Pointwise Convolution Unit, Variable Density Bound Block, Activation stationary, Sparsity, MobileNet-v1.

ABSTRACT

MobileNet introduced depthwise separable convolution, significantly reducing computations in comparison to standard convolution, making them fit for edge computing. Parameter-efficient models like MobileNet-v1 can be pruned to further reduce the number of parameters. In most of the pruning methods, pointwise convolutional parameters are pruned as they are large in number and depthwise convolutional parameters are kept dense. The efficacy of MobileNet pruning for edge deployment depends critically on a sparsity-aware execution environment. This paper focuses on implementing sparse aware pointwise convolution unit that leverage weight sparsity using Variable Density Bound Block (VDBB) technique. The activation stationary method is also incorporated, as it saves the power to fetch activations continuously. The proposed architecture is implemented on the FPGA platform using Vivado 2024.1. Implementing the sparse-aware architecture on the xc7z045ffv900-1 FPGA resulted in a 358mW reduction in power, a 50% reduction in DSPs, and a 24.06% reduction in LUTs compared to the sparse-unaware architecture. An engine from the proposed architecture was synthesized in ASIC using the TSMC 90 nm standard cell library, resulting in a 30.87% reduction in power and a 32.89% reduction in area when compared to a sparse unaware engine.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Convolution Neural Networks (CNNs) are crucial in modern signal processing, such as classification, denoising, and pattern recognition. CNN can learn relevant features directly from the data, reducing the need for manual feature engineering. The popular architectures among CNN are Dense CNN [1], AlexNet [2], VGG [3], and GoogLeNet [4]. These architectures have millions of parameters and billions of computations. Storing them at a single-precision floating-point demands Megabits of storage space. Performing millions of computations demands enormous power. More storage space and high-power consumption are not desirable for edge devices. The dense model parameters can be reduced by using methods such as quantization, pruning, and so forth. Quantization contributes to reduction in memory footprint as it reduces the precision of weights and activations.

Pruning decreases the quantity of parameters; pruning can be structured or unstructured contributing to block sparsity and tensor sparsity respectively. Also, there are lightweight models like MobileNet-v1 [5], MobileNet-v2 [6], ShuffleNet [7], and Squeeze Net [8] with a smaller number of parameters and computations but with a considerable amount of accuracy drop. MobileNet, ShuffleNet, and EfficientNet-Lite use Depthwise separable convolution, as the number of computations got reduced by $((1/N) + (1/D_k^2))$ compared to standard convolution, where N is number of kernels and Dk is the dimension of kernel. The question arises: can lightweight models like MobileNet be pruned? as it is less redundant. But certain pruning techniques for MobileNet-v1 yielded remarkable results. Automated Gradual Magnitude Pruning (GMP) algorithm [9] prunes the lowest magnitude weights, demonstrated that pruned dense MobileNet with 50% sparsity is better in accuracy than 0.75 MobileNet with a lesser number of parameters.

Soft Threshold Weight Reparameterization (STR) [10] achieved good accuracy for CNN in case of unstructured sparsity. STR achieved an accuracy of 68.35% using 101M FLOPS (FLoating point Operations Per Second) at 75.28% sparsity, where GMP achieved an accuracy of 67.7% using 163M FLOPS at 74.11% sparsity for MobileNet-v1. Deep Q-Learning for Pruning Deep Neural Networks (QLP) [11] method based on deep reinforcement learning achieved a good amount of accuracy 71.06% at 50% sparsity with an accuracy drop of 0.89% compared to the baseline MobileNet-v1. These results demonstrate that MobileNet-v1 can be pruned. It is interesting to note that the first layer and the depthwise convolutional layers are kept dense in GMP, QLP methods. STR made depthwise convolutional layers less sparse.

So, only Pointwise convolutional layers are pruned in these methods as it has 74.59% of parameters [5]. All these methods contribute to unstructured sparsity. Unstructured sparsity is advantageous while deploying the model on edge devices only if the designed hardware is sparse aware. So, this work focuses on implementing sparse-aware architecture for pointwise convolutions using the Variable Density Bound Block technique, which works for all sparsity levels without losing non-zero parameters. The activations are kept stationary until the dot product with all the kernels are computed, which reduces the power consumption to bring activations from the buffer to the processing units continuously.

II. POINTWISE CONVOLUTION

Pointwise convolution is also known as 1x1 convolutions. Usually, it is paired with depthwise convolution in MobileNet-v1, as depthwise convolution lacks channel mixing to learn complex hierarchical features [5]. Pointwise convolution is utilized in ResNet bottlenecks to expand and compress channels, in ShuffleNet for channel mixing, and in attention blocks for channel recalibration. So, pointwise convolution is used for channel mixing and for dimension adjustments. Pointwise convolution is quite like standard convolution but the only difference is its kernel spatial dimensions.

In pointwise convolution, the kernel dimension is 1x1xC where C is the number of channels. Pointwise convolution is nothing but the dot product of the kernel and the part of the activation layer as shown in Figure.1. Assuming an activation layer of dimension HxWxC and N number of kernels of dimension 1x1xC, then the pointwise convolution output dimension is HxWxN. N decreases the channel dimension if it is less than C and increases if it is greater than C. Mathematically, let activation layer be A $\in \mathbb{R}^{HxWxC}$ and kernel be W1 $\in \mathbb{R}^{1x1xC}$ then pointwise convolution output at a particular location (h, w) is

$$P(h,w) = \sum_{c'=1}^{C} A(h,w,c') \,.\, W_1(1,1,c') \quad (1)$$

Where

H= Number of rows of activation layer

W= Number of columns of activation layer

C= Number of channels

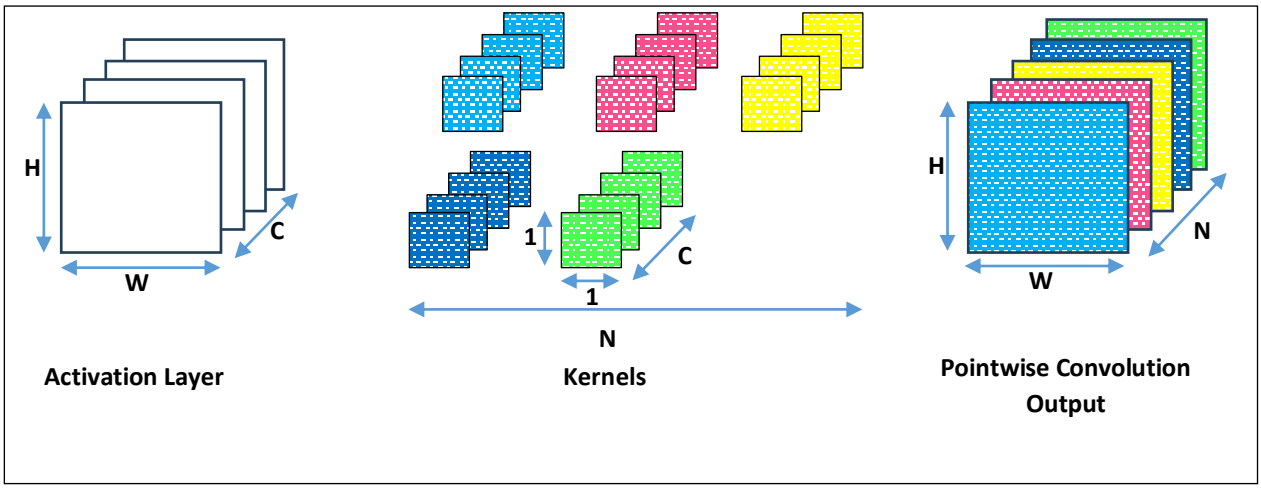


Figure 1: Pointwise Convolution.
Source: Authors, (2026).

III. PROPOSED ARCHITECTURE

Figure.2. shows the proposed architecture for sparse aware pointwise convolution unit. The proposed architecture supports parallelism by segregating feature maps into tiles [12]. The proposed architecture was designed for MobileNet-v1, but it can also be used for different models where pointwise convolutions are needed, just by redesigning buffers. MobileNet-v1 feature maps have channel sizes that are multiples of 32, with the exception of input image channels, as shown in Figure.3. So, each data word in the activation buffer contains 32 activations, which are taken in channel-wise direction to enable 32 channel unrolling. According to the position of non-zero weights, a vector of 1's and 0's was stored as a mask in the mask buffer as shown in Figure.4. As the weights are sparse, the weight buffer conserves memory by storing only non-zero weights. Entire activations along channel-wise at a particular location are fetched into the activation register from the activation buffer. For instance, if there are 64 channels, then two data words are fetched into the activation register. These activations in the register are kept stationary until all kernels are convolved with these activations.

The activation selector uses Variable Density Bound Block (VDBB) to choose activations corresponding to non-zero weights from the activation register using mask value. Because of VDBB, each engine has sixteen multipliers for 32-channel unrolling. The concept of VDBB is discussed in detail in section 4. Each engine takes one clock cycle if sparsity is greater than 50% and two clock cycles if sparsity is less than 50% for 32 channels. Partial results are accumulated until all channels at a specific spatial location within the feature map are processed.

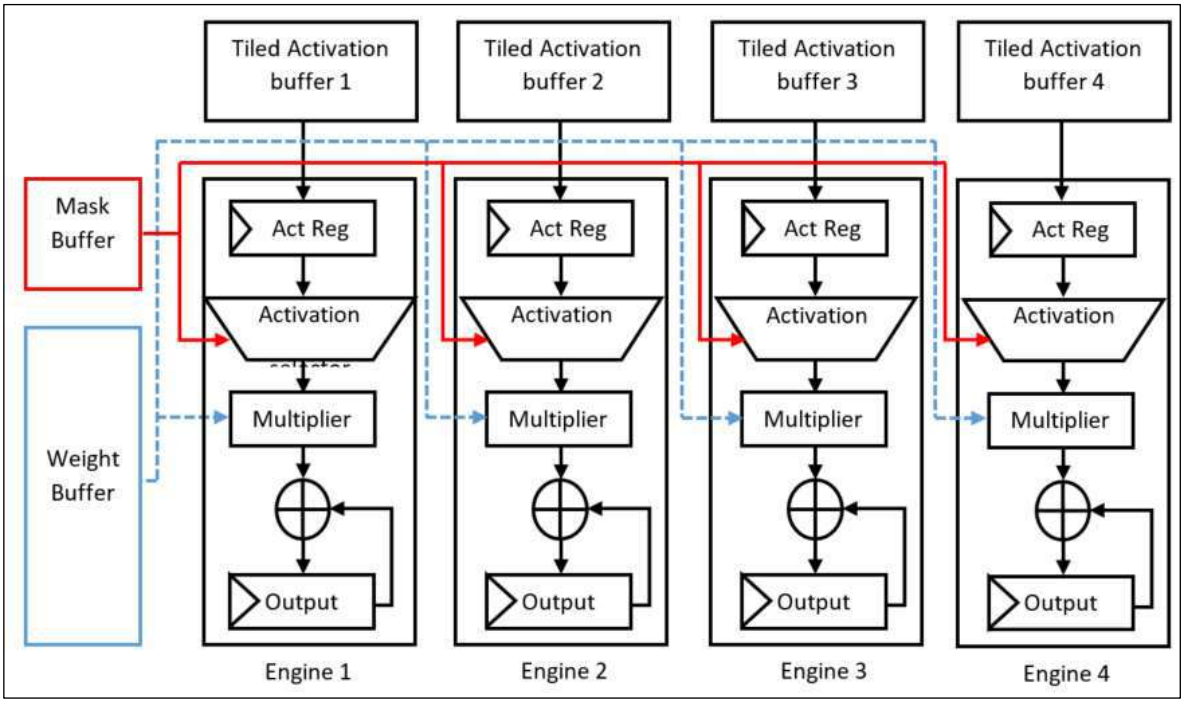


Figure 2: Proposed Architecture of Sparse-Aware Pointwise Convolution Unit.
Source: Authors, (2026).

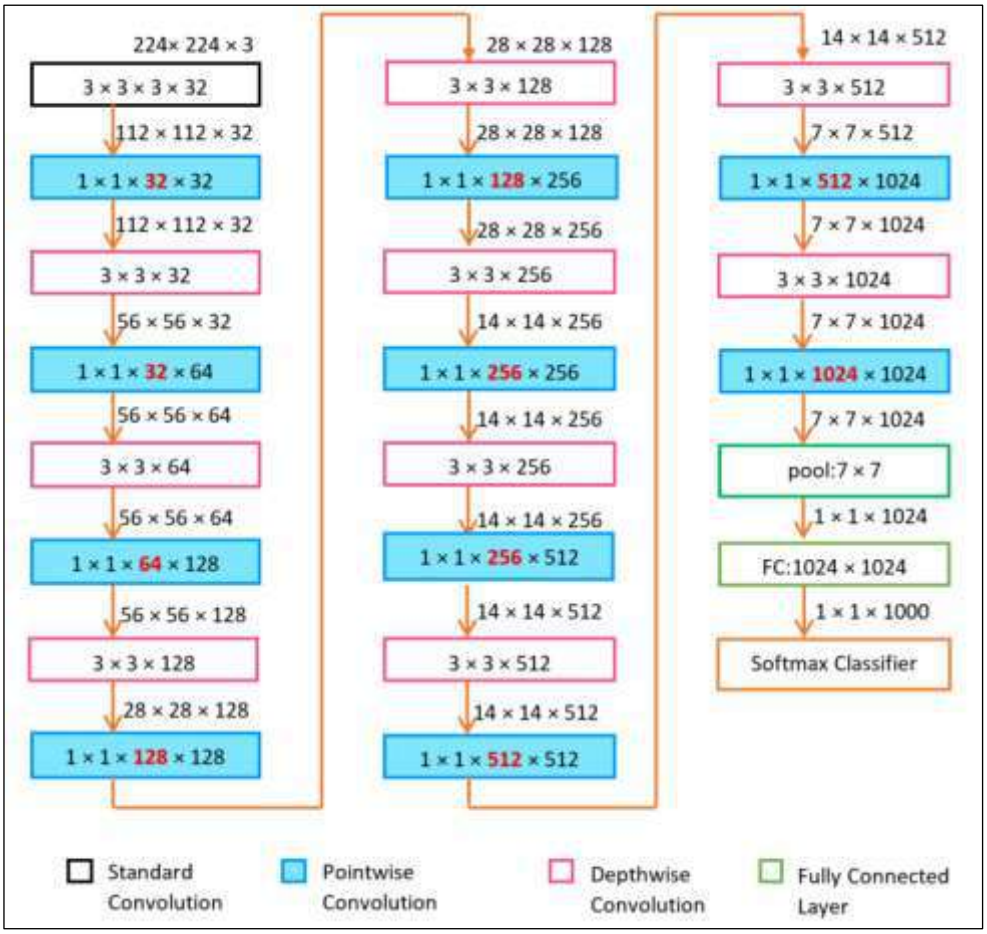


Figure 3: MobileNet-v1 Layer Information.
Source: Authors, (2026).

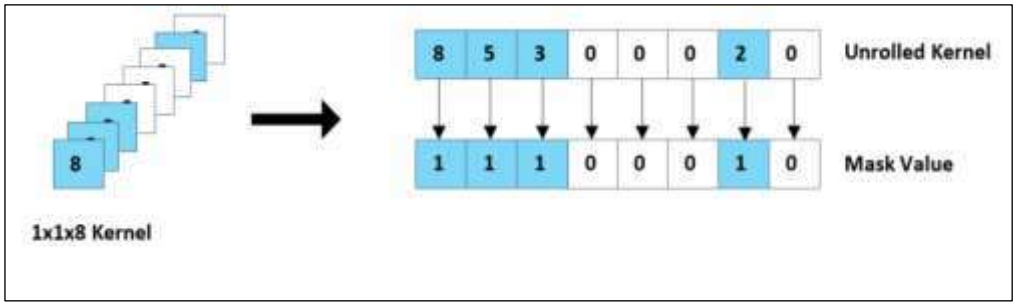


Figure 4: Mask generation.
Source: Authors, (2026).

IV.ACTIVATION SELECTOR

Sparsity is nothing but the existence of zeros in the weight tensor or in activation tensor. Pruning causes sparsity in weights, while ReLU causes sparsity in activations. Based on the distribution of non-zero elements in sparse tensors, sparsity can be Random, block sparse, density bound block as shown in Figure.5. If non-zero elements are distributed randomly, as illustrated in Figure. 5(a), this configuration is referred to as random sparse. Usually sparsity in Convolutional Neural Networks (CNNs) is random. This randomness creates an overhead in hardware while inferencing the model, as indexes of non-zero elements need to be stored specifically and the load is also unpredictable. In the context of block sparsity, where non-zero elements are organized in blocks, as illustrated in Figure. 5(b), fewer indexes are required compared to a random arrangement. However, imposing block sparsity on CNNs can lead to a decline in accuracy. In Density Bound Block (DBB) [13], the number of non-zero elements in a block are predefined, instead of forcing all the elements in a block to zero or fully dense as shown in Figure.5(c), which saves from accuracy drop.

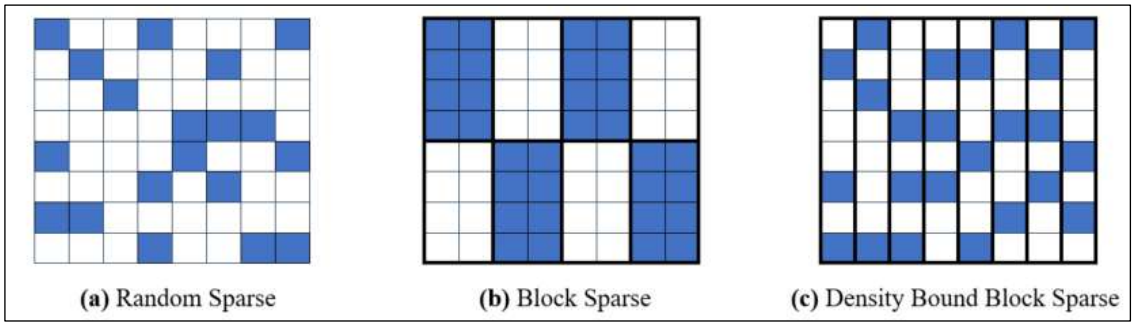


Figure 5: Types of Sparsity, filled boxes represents non-zero elements and empty boxes represent zeros.
Source: Authors, (2026).

DBB enables the use of Multiply and Accumulate (MAC) units in proportion to the level of sparsity. For instance, with 50% sparsity, only 50% of the usual MACs are needed, which significantly lowers hardware costs. But the problem is, sparsity is fixed at design time leading to loss of non-zero elements when sparsity is less than predefined sparsity. To overcome this issue, Variable Density Bound Block Technique (VDBB) [14] had been designed, which works at all sparsity levels. VDBB uses the same number of MACs as DBB. As the weights are randomly sparse, and the engine in the architecture processes 32 weights at once, this results in varying sparsity levels. To handle these varying sparsity levels, the proposed architecture uses VDBB as activation selector. Thus, no non-zero element is discarded during computation. The design for VDBB is taken from [15]. As shown in Figure. 6, sparsity less than 50% takes two clock cycles to compute, and sparsity greater than 50% takes one clock cycle to compute.

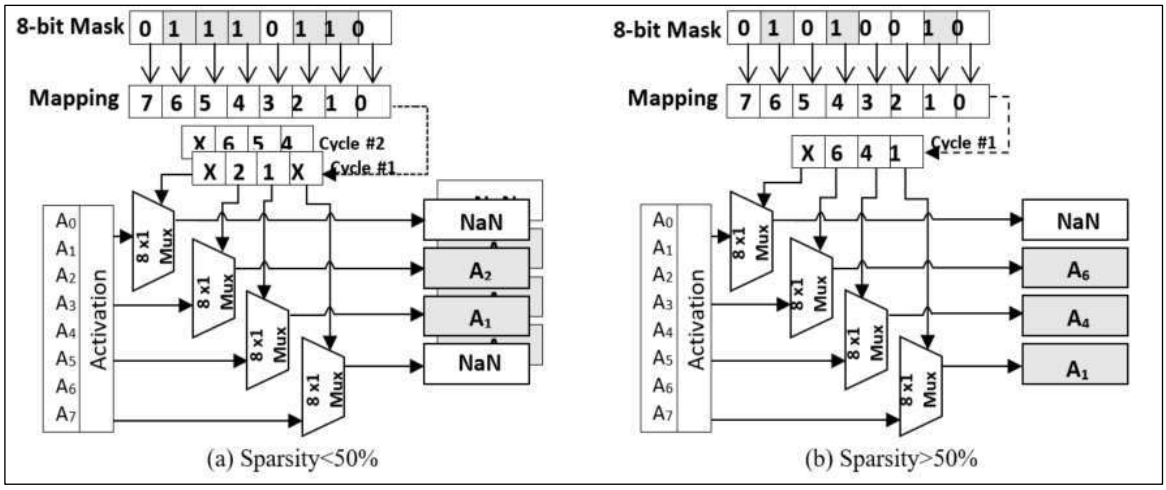
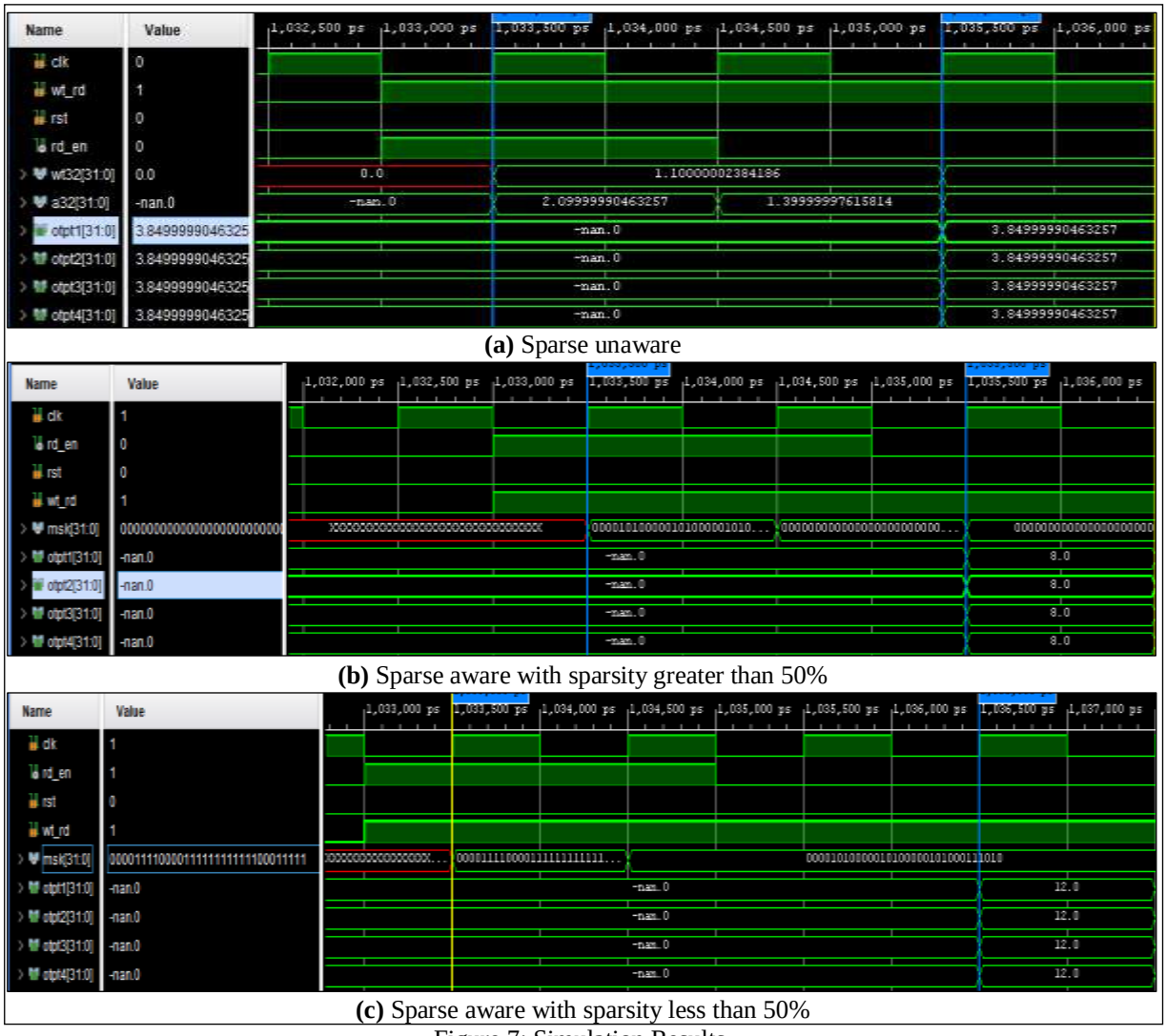


Figure 6: Activation Selector based on VDBB.
Source: Authors, (2026).

The mask value is used to determine the index values to choose activations corresponding to non-zero weights. Indexes are given to multiplexers, which select the corresponding activations. The indexes are organized in such a way that if sparsity is greater than 50%, then the indexes are accumulated in one clock cycle as shown in Figure. 6(a); if sparsity is less than 50%, then the indexes are organized in two clock cycles as shown in Figure.6(b).

V. RESULTS AND DISCUSSIONS

The proposed architecture had been implemented using Xilinx Vivado 2024.1 targeting xc7z045ffv900-1. Verilog HDL is used to design the architecture. Figure.7 shows the simulation traces for sparse-unaware and sparse-aware architectures. For the depth of 64, it is observed from Figure.7(a) and Figure.7(b) that in both sparse unaware and sparse aware (both data words exhibit >50% sparsity) simulation, two clock cycles are needed to get the final output, as two data words need to be processed. Figure. 7(c) presents a mixed-sparsity case in which one data word has sparsity below 50% (requiring two cycles) while the other exceeds 50% (requiring one cycle), resulting in a total of three clock cycles. These results indicate that the sparsity-aware implementation incurs an extra clock cycle whenever a processed data word's sparsity falls below the 50% threshold. Activation-stationary behavior is evident from the rd_en signal, as shown in Figure 7. The signal asserts only when activation words need to be fetched from the activation buffer according to the configured depth, thereby eliminating continuous buffer reads and reducing power consumption.



(a) Sparse unaware

(b) Sparse aware with sparsity greater than 50%

(c) Sparse aware with sparsity less than 50%

Figure 7: Simulation Results.
Source: Authors, (2026).

From Table.1, it is observed that the proposed architecture resulted in 50% reduction in DSPs, 24.06% reduction in LUTs, 8.96% reduction in BRAMs and 10.96% rise in Flip-flops compared to sparse unaware architecture. The results clearly demonstrate that the proposed architecture achieved good resource utilization, enabling efficient deployment on edge devices. It is also observed that the power was reduced by 358mW compared to sparse unaware architecture. ASIC synthesis of the sparse-aware and sparse-unaware engines was carried out in Cadence Genus and Innovus, targeting a 90 nm standard cell library. The core area of sparse aware architecture is 0.55980mm^2, whereas sparse unaware architecture occupies 0.83421mm^2. Although the number of MACs was reduced by 50% compared to the sparse-unaware engine, the inclusion of activation selector units resulted in an overall area reduction of 32.89%. The proposed engine consumes 0.91mW, whereas the sparse-unaware engine consumes 1.32mW as shown in Table 2. Figure.8 shows the layout of the sparse aware engine implemented using Innovus.

Table 1: Performance comparison of sparse unaware and sparse aware architectures implemented using FPGA xc7z100ffg900-1.

	Sparse Unaware architecture	Sparse aware architecture
# LUT	64180	48733
# LUTRAM	2736	2736
# FF	4467	5017
#BRAM	72.50	66
# DSP	256	128
Precision	FP32	FP32
Operating Frequency	100MHz	100MHz
Power(W)	2.231	1.873

Source: Authors, (2026).

Table 2: Performance comparison of ASIC implemented sparse unaware and sparse aware engine.

	Sparse Unaware architecture	Sparse aware architecture
Area(mm^2)	0.83421	0.55980
Power(mW)	1.32787	0.91793

Source: Authors, (2026).

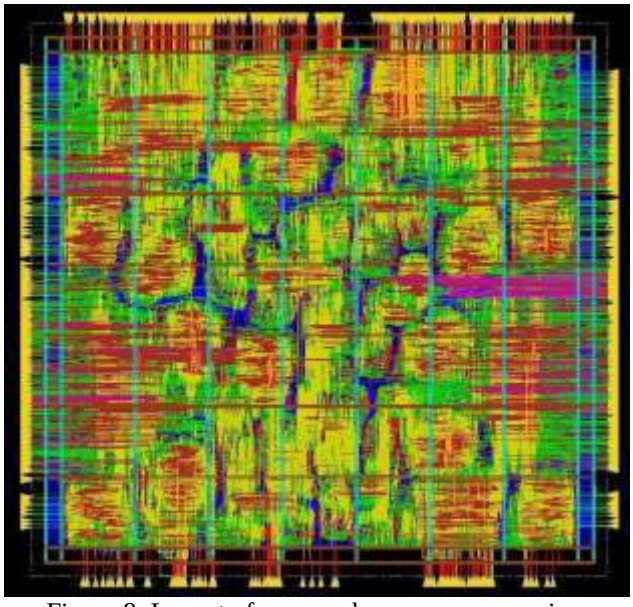


Figure 8: Layout of proposed sparse aware engine.
Source: Authors, (2026).

VI. CONCLUSION

Implementing Neural networks on edge devices demands efficient models and efficient hardware, as neural networks are computationally intensive. There are lightweight models like MobileNet, EfficientNet, SqueezeNet, and so on, with greatly reduced parameters. Even these lightweight models are further pruned using various techniques without much accuracy loss to fit in edge devices. These techniques prune only point wise layers, as they are huge in number. Pruning causes sparsity, this is advantageous only when the hardware is sparse aware. So, the sparse-aware pointwise convolution unit architecture was designed and implemented on FPGA, and an engine of this architecture was implemented in ASIC. In FPGA implementation, sparse aware architecture resulted in 50% reduction of DSPs and 358mW reduction in power compared to sparse unaware. Implementation of the engine in ASIC using 90nm resulted in 32.89% reduction in area and 30.87% reduction in power. These results emphasize the crucial role of sparse-aware design while deploying neural networks on edge devices.

VII.AUTHOR'S CONTRIBUTION

Conceptualization: Golajapu Jyothsna, Narayanam Balaji.
Methodology: Golajapu Jyothsna, Narayanam Balaji.
Investigation: Golajapu Jyothsna, Narayanam Balaji.
Discussion of results: Golajapu Jyothsna, Narayanam Balaji.
Writing – Original Draft: Golajapu Jyothsna, Narayanam Balaji.
Writing – Review and Editing: Golajapu Jyothsna, Narayanam Balaji.
Resources: Golajapu Jyothsna, Narayanam Balaji.
Supervision: Golajapu Jyothsna, Narayanam Balaji.
Approval of the final text: Golajapu Jyothsna, Narayanam Balaji

VIII. REFERENCES

[1] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” Jan. 2018, [Online]. Available: http://arxiv.org/abs/1608.06993

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Commun ACM, vol. 60, no. 6, pp. 84–90, Jun. 2017, doi: 10.1145/3065386.

[3] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Apr. 2015, [Online]. Available: http://arxiv.org/abs/1409.1556

[4] C. Szegedy et al., “Going Deeper with Convolutions,” Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.4842

[5] A. G. Howard et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” Apr. 2017, [Online]. Available: http://arxiv.org/abs/1704.04861

[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Dec. 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.

[7] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices,” Dec. 2017, [Online]. Available: http://arxiv.org/abs/1707.01083

[8] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size,” Nov. 2016, [Online]. Available: http://arxiv.org/abs/1602.07360

[9] M. Zhu and S. Gupta, “To prune, or not to prune: exploring the efficacy of pruning for model compression,” Nov. 2017, [Online]. Available: http://arxiv.org/abs/1710.01878

[10] A. Kusupati et al., “Soft Threshold Weight Reparameterization for Learnable Sparsity,” Jun. 2020, [Online]. Available: http://arxiv.org/abs/2002.03231

[11] E. Camci, M. Gupta, M. Wu, and J. Lin, “QLP: Deep Q-Learning for Pruning Deep Neural Networks,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 10, pp. 6488–6501, Oct. 2022, doi: 10.1109/TCSVT.2022.3167951.

[12] H. Hong, D. Choi, N. Kim, and H. Kim, “Mobile-X: Dedicated FPGA Implementation of the MobileNet Accelerator Optimizing Depthwise Separable Convolution,” IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 71, no. 11, pp. 4668–4672, 2024, doi: 10.1109/TCSII.2024.3440884.

[13] Z.-G. Liu, P. N. Whatmough, and M. Mattina, “Systolic Tensor Array: An Efficient Structured-Sparse GEMM Accelerator for Mobile CNN Inference,” May 2020, [Online]. Available: http://arxiv.org/abs/2005.08098

[14] Z.-G. Liu, P. N. Whatmough, and M. Mattina, “Sparse Systolic Tensor Array for Efficient CNN Hardware Acceleration,” Sep. 2020, [Online]. Available: http://arxiv.org/abs/2009.02381

[15] Q. Cheng et al., “A Low-Power Sparse Convolutional Neural Network Accelerator With Pre-Encoding Radix-4 Booth Multiplier,” IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 70, no. 6, pp. 2246–2250, Jun. 2023, doi: 10.1109/TCSII.2022.3231361.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 581-587. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2863

RESEARCH ARTICLE OPEN ACCESS

INTEGRATION OF MOST AND TOPSIS METHODS TO SUPPORT BUSINESS DEVELOPMENT STRATEGY FORMULATION IN THE STEEL PIPE INDUSTRY OF SURABAYA, INDONESIA

Rachmat Yustiawan Hadi[1] and Mohamad Ferdaus Noor Aulady*[2]

[1]Department of Industrial Engineering, Faculty of Industrial Technology, Institut Teknologi Adhi Tama Surabaya, Surabaya, Indonesia.
[2]Department of Civil Engineering, Faculty of Civil Engineering and Planning, Institut Teknologi Adhi Tama Surabaya, Surabaya, Indonesia.

[1]https://orcid.org/009-0009-2699-1977, [2]https://orcid.org/0000-0001-9079-9458

Email: *mohamadaulady@itats.ac.id, rachmatyustiawan@gmail.com

ARTICLE INFO

Article History
Received: October 28, 2025
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
MOST,
AHP,
TOPSIS,
Business strategy,
Steel pipe industry.

ABSTRACT

This study develops an integrated decision-making framework to formulate business development strategies for the steel pipe manufacturing industry in Surabaya, Indonesia. The research combines the MOST and TOPSIS methods, supported by the Analytical Hierarchy Process (AHP), to systematically identify, evaluate, and prioritize strategic alternatives. Primary data were collected through expert interviews, focus group discussions, and questionnaires, while secondary data were obtained from industrial reports and publications. The analysis identifies five strategic alternatives derived from the MOST framework and evaluates them using AHP–TOPSIS integration. The results show that investment in production technology and automation ranks as the highest-priority strategy, followed by the development of export-oriented marketing channels and enhancement of quality management systems. Sensitivity analysis confirms the robustness of these rankings. The integrated MOST–AHP–TOPSIS model proves effective in aligning organizational missions with quantitative decision-making, resulting in more objective and evidence-based strategy formulation. This framework provides a practical tool for industrial managers to enhance competitiveness, operational efficiency, and strategic adaptability within the steel pipe sector.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The steel pipe industry plays a strategic role in supporting various sectors such as construction, manufacturing, oil and gas, and national infrastructure development. In Indonesia, especially in Surabaya as one of the major industrial hubs, the demand for steel pipes continues to grow in line with rapid economic expansion and infrastructure projects. However, steel pipe manufacturers face multiple challenges, including intense domestic and global competition [1], raw material price volatility [2], technological shifts in manufacturing [3], and government regulations on trade and environmental standards [4]. Although several companies in Surabaya show positive growth in production capacity and sales, many still encounter difficulties in maintaining competitiveness, improving operational efficiency, and achieving sustainable profitability. These conditions highlight the urgent need for a well-formulated and adaptive business development strategy that aligns internal capabilities with external opportunities in the dynamic steel pipe market [5]. Previous studies commonly apply strategy formulation frameworks such as SWOT–AHP [6], [7]or AHP–TOPSIS [8], [9] to evaluate competitive priorities and strategic decision alternatives. However, limited research has explored the integration of the MOST (Mission, Objectives, Strategy, Tactics) framework with the TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) method in industrial business strategy formulation. The MOST framework provides a structured qualitative approach to aligning an organization's vision and mission with measurable objectives and actionable tactics [10].

In contrast, the TOPSIS method offers a quantitative multi-criteria decision-making model that ranks strategic alternatives based on their closeness to the ideal solution [11]. Integrating the MOST and TOPSIS methods enables companies to combine qualitative

strategic alignment with quantitative decision-making, resulting in strategies that are both analytically sound and practically applicable. This integration helps organizations link long-term missions with data-driven evaluations to achieve more rational and objective business priorities. Accordingly, this study focuses on examining how the integration of the MOST and TOPSIS methods can support the formulation of optimal business development strategies for steel pipe manufacturing companies in Surabaya. To achieve this aim, the research establishes several objectives. First, it identifies internal and external factors that influence business development in Surabaya's steel pipe industry through the application of MOST and PEST analyses. Second, it designs strategic alternatives that align with the company's mission, objectives, and tactical priorities to ensure coherence between strategic vision and operational execution. Third, the study determines the relative importance or weight of each strategic criterion using the Analytical Hierarchy Process (AHP), enabling a structured and quantitative evaluation of priorities.

Finally, it employs the TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) method to rank and prioritize the most effective business development strategy alternatives. Collectively, these stages establish an integrated decision-making framework that combines qualitative insight with quantitative analysis to formulate optimal and actionable business strategies for the steel pipe industry in Surabaya. This study assumes that integrating qualitative and quantitative approaches enhances the robustness of strategic decision-making. The research focuses on steel pipe manufacturers in Surabaya, using primary data from interviews, focus group discussions (FGDs), and questionnaires with industry stakeholders, supported by secondary data from industrial reports, publications, and company documents. The methodology combines qualitative analysis (MOST and PEST) to identify strategic dimensions and quantitative analysis (AHP–TOPSIS) to calculate and rank strategy alternatives. Data are processed using Super Decisions software for AHP weighting and Microsoft Excel for TOPSIS analysis. This research contributes theoretically by expanding the literature on integrated decision-making models in strategic management and practically by providing a systematic, data-driven framework for business development. The model can serve as a reference for steel pipe industries in Surabaya to enhance competitiveness, improve operational efficiency, and achieve sustainable industrial growth.

II. THEORETICAL REFERENCE

A business strategy serves as a structured plan designed to guide organizations in achieving their long-term goals through coordinated decisions, efficient resource allocation, and sustained competitive advantage. It functions as an effective managerial instrument to enhance organizational performance and competitive positioning [12]. Business strategy, also referred to as competitive strategy, focuses on improving the market position of products or services within specific industry segments [13]. An effective business strategy aligns internal capabilities with external environmental factors to ensure continuous adaptability in a dynamic industrial context.

II.1 MOST METHOD

The MOST method (Mission, Objectives, Strategy, and Tactics) provides a systematic framework for aligning organizational vision and mission with specific, measurable business objectives. It facilitates the logical linkage between long-term aspirations and short-term operational activities. The MOST framework encourages management to periodically review strategic coherence among its four components Mission, Objectives, Strategy, and Tactics ensuring that each operational activity supports organizational goal [14]. In practice, MOST enables decision-makers to translate vision statements into actionable strategies that drive competitive growth and sustainable business outcomes.

II.2 PEST ANALYSIS

Complementing MOST, the PEST analysis (Political, Economic, Social, and Technological) evaluates macro-environmental factors influencing corporate strategy. It assists organizations in identifying opportunities and anticipating external risks [15]. The political dimension encompasses government regulations, fiscal policy, and trade stability, while the economic aspect relates to market dynamics such as inflation, interest rates, and raw material prices. The social dimension focuses on demographic and cultural factors, whereas the technological aspect assesses innovation trends, digital transformation, and process automation all of which significantly affect the steel pipe industry's competitiveness in Indonesia.

II.3 AHP

The Analytical Hierarchy Process (AHP) is a structured multi-criteria decision-making method that uses pairwise comparisons to assign relative weights to different criteria. It decomposes complex problems into hierarchical structures, allowing quantitative prioritization of qualitative judgments. AHP has been widely adopted for resource allocation, supplier selection, and business strategy formulation [16]. In this research, AHP supports the evaluation of strategic alternatives identified through MOST and PEST by assigning consistent and measurable importance weights based on expert judgment.

II.3.1 The Geometric Mean

The geometric mean is applied to aggregate multiple expert evaluations into a single representative value while maintaining the reciprocal property of the original comparison matrix. This mathematical approach ensures consistency in collective decision-making when multiple respondents provide assessments independently. The geometric mean GM is computed as:

The geometric mean

$$G = \sqrt[n]{X_1 \times X_2 \times X_3 \ldots} \quad (1)$$

II.4 TOPSIS METHOD

The TOPSIS method (Technique for Order Preference by Similarity to Ideal Solution) is a multi-criteria decision-making approach based on ranking alternatives according to their geometric distances from the positive ideal solution and the negative ideal solution [17]. The best alternative is the one with the shortest distance to the positive ideal and the farthest from the negative ideal [18]. TOPSIS has been widely utilized in industrial decision support systems due to its computational efficiency, simplicity, and strong logical foundation. It has been applied in manufacturing strategy formulation, supplier evaluation, and investment prioritization [19]. Integrating MOST, AHP, and TOPSIS creates a balanced framework that combines qualitative strategic alignment with quantitative evaluation and ranking. MOST defines the strategic intent, AHP determines the importance of key criteria through expert-based weighting, and TOPSIS provides an objective ranking of strategic alternatives. This integration has been recognized in recent studies [20], [21], [22] as an effective hybrid approach for improving decision accuracy and ensuring that strategic decisions remain consistent with organizational mission and environmental dynamics.

III. MATERIALS AND METHODS

This research is conducted to develop an integrated decision-making framework for formulating business development strategies in the steel pipe manufacturing industry located in Surabaya, Indonesia. The study integrates the MOST and TOPSIS methods to produce strategic alternatives that are both analytically sound and practically applicable. The AHP is also employed to determine the relative importance of each criterion, enabling a structured weighting process that supports the quantitative assessment of strategies. The study focuses on medium- and large-scale steel pipe manufacturing companies operating in the Surabaya industrial corridor, which represent a significant segment of Indonesia's downstream metal sector. The research emphasizes strategic management aspects such as competitiveness, business positioning, and operational effectiveness rather than the technical aspects of steel production. The main limitation of this study lies in the scope of available managerial data, the use of subjective judgment in expert evaluations, and potential differences in perception among respondents during the AHP weighting process. The materials used in this research consist of both primary and secondary data.

Primary data are obtained through structured interviews, focus group discussions (FGDs), and questionnaires distributed to managerial-level respondents and industry experts, including production, marketing, and quality assurance managers with more than five years of experience in the steel pipe industry. Secondary data are gathered from company annual reports, official publications of the Indonesian Ministry of Industry, trade association bulletins, and relevant academic literature. The questionnaires are developed based on the MOST framework and designed using a five-point Likert scale to capture qualitative assessments, while the AHP stage uses pairwise comparison matrices to quantify expert preferences. Ten respondents are selected using purposive sampling to ensure representation from different functional areas. The research population refers to a major steel pipe manufacturing company in Surabaya that meets the criteria of operational continuity, integrated production facilities, and active participation in national industrial networks. This company represents a suitable and relevant case for studying business development strategies in Indonesia's steel pipe sector, as it embodies the competitive dynamics, market challenges, and organizational structures typical of the industry.

Respondents are selected through judgmental sampling, targeting managerial personnel directly involved in strategic planning and business development to ensure that the collected data accurately reflect managerial perspectives and decision-making practices. The research procedures begin with identifying strategic factors through the integration of MOST and PEST analyses. The results are used to construct hierarchical decision structures for AHP weighting. Pairwise comparison matrices are then developed, and consistency ratios (CR) are tested to ensure all CR values are below 0.1, indicating reliable expert judgment. The resulting AHP weights serve as input for the TOPSIS method. Data processing represents the practical implementation of the theoretical model by combining the outputs of MOST, AHP, and TOPSIS into a single computational framework. The process involves constructing normalized decision matrices, determining criterion weights, calculating distances to ideal and anti-ideal solutions, and ranking strategy alternatives based on the calculated closeness coefficients.

The final ranking results are verified through expert review to ensure consistency with strategic expectations. The choice of methods is justified by their complementary strengths: MOST provides strategic alignment with organizational objectives, AHP offers a systematic weighting mechanism, and TOPSIS ensures quantitative prioritization for rational and evidence-based decision-making. This integration enhances the credibility and applicability of the results by combining qualitative insight with quantitative rigor. All participants involved in the study provide informed consent before data collection, and confidentiality is maintained by anonymizing respondent identities and company-related information. Ethical standards are observed throughout the research process to ensure data integrity and respect for participant privacy. Overall, this methodological framework establishes a transparent, replicable, and reliable process for identifying, evaluating, and prioritizing business development strategies in the steel pipe manufacturing industry of Surabaya, Indonesia, through the integration of MOST, AHP, and TOPSIS approaches.

IV. RESULTS AND DISCUSSIONS

The integration of MOST, AHP, and TOPSIS provides a comprehensive analytical framework for identifying, evaluating, and prioritizing business development strategies for the steel pipe manufacturing industry in Surabaya, Indonesia. The analysis begins with identifying internal and external strategic factors using the MOST and PEST frameworks, determining the criterion weights through AHP, and ranking the alternative strategies using TOPSIS. The application of the MOST framework allows the company's mission, objectives, strategies, and tactics to be systematically aligned. The mission emphasizes strengthening competitiveness through technological innovation, operational efficiency, and sustainable business practices. The objectives include expanding domestic and international markets, optimizing production efficiency, and enhancing value-added capabilities. Based on these, five strategic alternatives are formulated:

Table 1: Strategic Alternatives Identified Through MOST Framework.

Code	Strategic Alternative	Strategic Focus
A1	Product diversification	Market and product innovation
A2	Investment in production technology and automation	Efficiency and productivity
A3	Development of export-oriented marketing channels	Market expansion
A4	Enhancement of quality management systems	Product reliability and compliance
A5	Digital transformation in supply chain and customer relations	Integration and digital resilience

Source: Authors, (2026).

The PEST analysis identifies external environmental factors influencing these strategic options. Politically, the Indonesian government's industrial downstreaming programs and infrastructure development projects increase market potential. Economically, global steel price fluctuations and exchange rate volatility influence cost structures. Socially, urbanization and industrial growth sustain long-term demand for construction and piping materials. Technologically, advances in automation, IoT-based monitoring, and digital manufacturing are reshaping competitiveness. Through AHP, experts evaluate the importance of strategic criteria derived from the MOST–PEST analysis. The pairwise comparison results yield a consistency ratio (CR) below 0.1, confirming judgment reliability. The highest-weighted criteria include market expansion potential (0.318), technological innovation (0.262), and operational efficiency (0.204), followed by quality improvement (0.132) and resource sustainability (0.084).

Table 2: AHP Results – Strategic Criteria and Weights.

Criterion	Weight	Priority Rank
Market expansion potential	0.318	1
Technological innovation	0.262	2
Operational efficiency	0.204	3
Quality improvement	0.132	4
Resource sustainability	0.084	5
Total	**1.000**	

Source: Authors, (2026).



Figure 1: AHP Weight Distribution of Strategic Criteria.
Source: Authors, (2026).

The AHP results are used as input for the TOPSIS analysis, which measures each strategic alternative's relative closeness to the ideal solution. The normalized decision matrix and weighted scores are processed.

Table 3: TOPSIS Ranking Results.

Code	Strategic Alternative	D_i^+	D_i^-	Closeness Coefficient (C_i)	Rank
A1	Product diversification	0.128	0.221	0.631	5
A2	Investment in production technology and automation	0.072	0.384	0.842	1
A3	Development of export-oriented marketing channels	0.089	0.327	0.786	2
A4	Enhancement of quality management systems	0.103	0.259	0.721	3
A5	Digital transformation in supply chain and customer relations	0.117	0.235	0.668	4

Source: Authors, (2026).



Figure 2: OPSIS Ranking of Strategic Alternatives.
Source: Authors, (2026).

The ranking shows that investment in production technology and automation (A2) achieves the highest closeness coefficient ($C_i = 0.842$), followed by development of export-oriented marketing channels (A3) and enhancement of quality management systems (A4). These results emphasize that the most critical strategic direction lies in leveraging technological innovation to improve efficiency and cost competitiveness while expanding to international markets. A sensitivity analysis is conducted to test the robustness of the ranking by varying AHP weights by ±10%. The ranking order remains stable, confirming that the results are not highly sensitive to small fluctuations in criteria weighting.

Table 4: Sensitivity Analysis of TOPSIS Results.

Scenario	Key Weight Adjustment	Top Strategy (Rank 1)	Rank Consistency
Base Case	Original weights	A2 (Technology & Automation)	-
Scenario 1	+10% Market expansion	A2	Stable
Scenario 2	+10% Technological innovation	A2	Stable
Scenario 3	+10% Operational efficiency	A2	Stable
Scenario 4	+10% Quality improvement	A2	Stable

Source: Authors, (2026).

From a managerial perspective, the prioritization indicates that technological advancement and automation are central to improving productivity and market responsiveness. Automation supports consistent quality control, minimizes defect rates, and enhances process efficiency. The second-ranked strategy, developing export-oriented marketing channels, aligns with Indonesia's trade liberalization and regional integration within ASEAN markets. Meanwhile, quality management improvement, though ranked third, remains essential for ensuring product reliability and achieving international certification such as ISO 9001 and API standards. The integrated application of MOST, AHP, and TOPSIS demonstrates strong methodological coherence. MOST provides strategic direction and goal alignment; AHP delivers structured weighting and quantification of managerial priorities; and TOPSIS translates these into objective rankings. This synergy produces more transparent and evidence-based decision outcomes than using a single qualitative or quantitative method alone.
The findings are consistent with previous research showing that AHP–TOPSIS integration enhances accuracy and rationality in industrial strategic decision-making [23], [24]. The inclusion of MOST in this study strengthens the theoretical contribution by ensuring that the strategy formulation process remains anchored to organizational mission and objectives. Practically, these results provide a strategic roadmap for steel pipe manufacturers in Surabaya to enhance competitiveness. Investment in automation technologies can optimize production efficiency and reduce operational variability, while export market development enables diversification and resilience against domestic market fluctuations. Continuous improvement in quality management and digital transformation can further support long-term sustainability and industry leadership. Overall, the integrated model successfully identifies and prioritizes business development strategies in the steel pipe manufacturing sector. The findings validate that the combination of MOST, AHP, and TOPSIS enhances both analytical depth and practical applicability, serving as a reliable reference for decision-makers in industrial strategic planning.

V. CONCLUSIONS

This research successfully develops an integrated decision-making framework for business development strategy formulation in the steel pipe manufacturing industry of Surabaya, Indonesia, by combining the MOST, AHP, and TOPSIS methods. The integration of these three approaches ensures that qualitative strategic alignment and quantitative prioritization are addressed simultaneously, resulting in more objective and data-driven strategic outcomes. The results of the analysis show that among the identified strategic alternatives, investment in production technology and automation holds the highest priority, followed by development of export-oriented marketing channels and enhancement of quality management systems. These findings indicate that technological advancement and market expansion are key factors in strengthening competitiveness and achieving sustainable growth in the steel pipe industry. The use of AHP enables accurate determination of the relative importance of strategic criteria such as market potential, technological innovation, operational efficiency, and quality improvement, while the TOPSIS method effectively ranks alternatives based on their closeness to the ideal solution.

The consistency ratio of AHP weighting (CR < 0.1) confirms the reliability of expert judgments, and the sensitivity analysis validates the robustness of the final rankings. Overall, the integration of MOST, AHP, and TOPSIS proves to be an effective and replicable framework for strategic decision-making in industrial management. It allows companies to align mission-driven objectives with quantitative evidence, offering a structured roadmap for enhancing competitiveness, operational efficiency, and long-term business sustainability. This model can serve as a valuable reference for other manufacturing sectors seeking systematic and evidence-based approaches to strategic planning.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Rachmat Yustiawan Hadi and Mohammad Ferdaus Noor Aulady.
Methodology: Rachmat Yustiawan Hadi and Mohammad Ferdaus Noor Aulady.
Investigation: Rachmat Yustiawan Hadi and Mohammad Ferdaus Noor Aulady.
Discussion of results: Rachmat Yustiawan Hadi and Mohammad Ferdaus Noor Aulady.
Writing – Original Draft: Rachmat Yustiawan Hadi.
Writing – Review and Editing: Rachmat Yustiawan Hadi and Mohammad Ferdaus Noor Aulady.
Resources: Mohamad Ferdaus Noor Aulady
Supervision: Mohammad Ferdaus Noor Aulady.
Approval of the final text: Rachmat Yustiawan Hadi and Mohammad Ferdaus Noor Aulady.

VII. ACKNOWLEDGMENTS

The author would like to express sincere gratitude to the Directorate General of Higher Education, Research, and Technology, Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia for the support provided through the Publication Assistance Program in Reputable Journals Year 2025 under contract number 0488/C/DT.06.01/2025. This assistance has significantly contributed to the successful publication of this article in a reputable international journal.

VIII. REFERENCES

[1] School of Business and Management, Institut Teknologi Bandung and R. Kristanto, "Proposed Marketing Strategy to Increase Revenue for PT. Pipa Tanpa Kampuh Indonesia in Seamless Steel Pipe Industry," Int. J. Curr. Sci. Res. Rev., vol. 05, no. 07, July 2022, doi: 10.47191/ijcsrr/V5-i7-10.

[2] School of Business Management, Bandung Institute of Technology, Bandung, Indonesia, N. S. Maharani, R. Aswin Rahadi, and School of Business Management, Bandung Institute of Technology, Bandung, Indonesia, "Leveraging Z-Score and
Financial Ratio as Early Warning System to Mitigate Supply Chain Disruption at PT Gunung Raja Paksi TBK," Int. J. Curr. Sci. Res. Rev., vol. 07, no. 09, Sept. 2024, doi: 10.47191/ijcsrr/V7-i9-07.

[3] S. Harianto, M. L. Singgih, I. M. L. Batan, I. Aulia, W. Sulistiyowati, and A. A. Permatasari, "Sustainable Innovation in HFW Steel Pipe Manufacturing: A Synergistic Approach of Lean Tools and Material Circularity," J. Posthumanism, vol. 5, no. 6, pp. 2026–2043, June 2025, doi: 10.63332/joph.v5i6.2310.

[4] M. F. N. Aulia and P. Lestari, "Gate-to-gate approach in life cycle assessment of steel pipe products," E3S Web Conf., vol. 485, p. 01008, 2024, doi: 10.1051/e3sconf/202448501008.

[5] Ural Federal University named after the first President of Russia B.N. Yeltsin et al., "Dynamic Assessment of the Company's Competitiveness, Taking into Account the Implementation of Its Innovative Development Strategy," J. Appl. Econ. Res., vol. 19, no. 4, pp. 512–542, 2020, doi: 10.15826/vestnik.2020.19.4.024.

[6] M. Abdel-Basset, M. Mohamed, and F. Smarandache, "An Extension of Neutrosophic AHP–SWOT Analysis for Strategic Planning and Decision-Making," Symmetry, vol. 10, no. 4, p. 116, Apr. 2018, doi: 10.3390/sym10040116.

[7] A. Görener, K. Toker, and K. Uluçay, "Application of Combined SWOT and AHP: A Case Study for a Manufacturing Firm," Procedia - Soc. Behav. Sci., vol. 58, pp. 1525–1534, Oct. 2012, doi: 10.1016/j.sbspro.2012.09.1139.

[8] A. Papapostolou, C. Karakosta, G. Apostolidis, and H. Doukas, "An AHP-SWOT-Fuzzy TOPSIS Approach for Achieving a Cross-Border RES Cooperation," Sustainability, vol. 12, no. 7, p. 2886, Apr. 2020, doi: 10.3390/su12072886.

[9] S. Cevik Onar, B. Oztaysi, and C. Kahraman, "Strategic Decision Selection Using Hesitant fuzzy TOPSIS and Interval Type-2 Fuzzy AHP: A case study:," Int. J. Comput. Intell. Syst., vol. 7, no. 5, p. 1002, 2014, doi: 10.1080/18756891.2014.964011.

[10] D. Globocnik, R. Faullant, and Z. Parastuty, "Bridging strategic planning and business model management – A formal control framework to manage business model portfolios and dynamics," Eur. Manag. J., vol. 38, no. 2, pp. 231–243, Apr. 2020, doi: 10.1016/j.emj.2019.08.005.

[11] S. Chaube et al., "An Overview of Multi-Criteria Decision Analysis and the Applications of AHP and TOPSIS Methods," Int. J. Math. Eng. Manag. Sci., vol. 9, no. 3, pp. 581–615, June 2024, doi: 10.33889/IJMEMS.2024.9.3.030.

[12] "Formulation of Business Strategies to Improve Business Performance by SWOT and SQSPM Approach in Era Pandemic: A Study on Culinary MSMEs," Qual. - Access Success, vol. 23, no. 188, Jan. 2022, doi: 10.47750/QAS/23.188.07.

[13] S. Sayudin, "Membentuk Strategi Bisnis Yang Tangguh Dalam Era Manajemen Yang Berubah [Formulating a Resilient Business Strategy in the Era of Changing Management]," J. Multidisiplin Indones., vol. 2, no. 7, pp. 1566–1572, July 2023, doi: 10.58344/jmi.v2i7.315.

[14] M. Iqbal, Md. G. Mahmood, and Md. R. Islam, "An Ergonomic Analysis of Working Posture of Selected Brick Kiln Workers using Rapid Upper Limb Assessment (RULA) and Rapid Entire Body Assessment(REBA) Method," in Proceedings of the International Conference on Industrial Engineering and Operations Management, Dhaka, Bangladesh: IEOM Society International, Dec. 2023. doi: 10.46254/BA06.20230027.

[15] Nadya Safitri and Rully Pramudita, "Perencanaan Strategis Sistem Informasi Menggunakan Metode PEST dan MOST [Strategic Information System Planning Using the PEST and MOST Methods]," Inf. Manag. Educ. Prof., vol. 3, no. 2, pp. 151–160, June 2019.

[16] J. Khazaii, “Analytical Hierarchy Process (AHP),” in Advanced Decision Making for HVAC Engineers, Cham: Springer International Publishing, 2016, pp. 73–85. doi: 10.1007/978-3-319-33328-1_9.

[17] P. Chen, “Effects of the entropy weight on TOPSIS,” Expert Syst. Appl., vol. 168, p. 114186, Apr. 2021, doi: 10.1016/j.eswa.2020.114186.

[18] M. Behzadian, S. Khanmohammadi Otaghsara, M. Yazdani, and J. Ignatius, “A state-of the-art survey of TOPSIS applications,” Expert Syst. Appl., vol. 39, no. 17, pp. 13051–13069, Dec. 2012, doi: 10.1016/j.eswa.2012.05.056.
[19] V. Pandey, Komal, and H. Dincer, “A review on TOPSIS method and its extensions for different applications with recent development,” Soft Comput., vol. 27, no. 23, pp. 18011–18039, Dec. 2023, doi: 10.1007/s00500-023-09011-0.

[20] S. Chaube et al., “An Overview of Multi-Criteria Decision Analysis and the Applications of AHP and TOPSIS Methods,” Int. J. Math. Eng. Manag. Sci., vol. 9, no. 3, pp. 581–615, June 2024, doi: 10.33889/IJMEMS.2024.9.3.030.

[21] Istanbul Technical University et al., “INTEGRATION OF ANALYTIC HIERARCHY PROCESS WITH OTHER MCDM METHODS: A LITERATURE REVIEW,” presented at the The International Symposium on the Analytic Hierarchy Process, Dec. 2022. doi: 10.13033/isahp.y2022.046.

[22] M. Hermansyah, M. Mujiono, F. Fatimatuzzahra, K. Dedes, and M. F. Firdausi, “Integration of AHP and TOPSIS Methods in Decision Making Models to Identify High Achieving Students,” J. Inform. Dev., vol. 3, no. 2, pp. 7–19, Apr. 2025, doi: 10.30741/jid.v3i2.1547.

[23] A. N. Alish Nazarov, “STRATEGIC MARKET ENTRY DECISION-MAKING IN AFRICA’S SATELLITE SECTOR: A COMPARATIVE ANALYSIS USING TOPSIS AND AHP METHODS,” PAHTEI-Procedings Azerbaijan High Tech. Educ. Inst., vol. 44, no. 09, pp. 85–94, Sept. 2024, doi: 10.36962/PAHTEI44092024-10.

[24] C.-H. Chen, “A Novel Multi-Criteria Decision-Making Model for Building Material Supplier Selection Based on Entropy-AHP Weighted TOPSIS,” Entropy, vol. 22, no. 2, p. 259, Feb. 2020, doi: 10.3390/e22020259.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 588-601. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2880

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

PERFORMANCE ANALYSIS OF EMOTION RECOGNITION USING YOLO

Varsha Patil*[1], Varun Chitale[2], Vijaya Pawar[3], Megha Yannwar[4]

[1, 2]Department of Electronics and Telecommunication Engg, AISSMS IOIT, Kennedy Road, 411001, Pune, India
[3]Department of Electronics and Telecommunication Engg, BVCOEW, Pune, India
[4]Department of Electronics and Telecommunication Engg, COEP Technological University, Pune, India

[1]https://orcid.org/0000-0003-4399-6410, [2]https://orcid.org/0009-0007-5568-6926
[3]https://orcid.org/0000-0002-9897-641X, [4]https://orcid.org/0009-0004-6673-6683

Email:*varshapatil101@gmail.com

ARTICLE INFO

Article History
Received: November 1, 2025
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Emotion recognition,
YOLO, FER2013,
Explainable AI,
FER,
Affective computing,
Object detection,
Emotions,
HCI,
Emotion intensity,
Temporal emotion tracking,
Mean Average Precision (mAP).

ABSTRACT

This paper presents Facial Emotion Recognition (FER) and Temporal emotion tracking System built using the YOLOv8 architecture. Proposed System not only recognises seven distinct emotions (angry, disgust, fear, happy, neutral, sad, and surprise) but maps emotion intensities. We trained our model using the FER2013 dataset, converting traditional classification data to an object detection format to leverage YOLOv8's capabilities. Unlike conventional emotion recognition approaches that simply categorise expressions, our system introduces three advanced analysis components: emotion intensity quantification that measures each emotion on a 0-100% scale, explainable AI visualisation using Grad-CAM that reveals which facial regions influence the model's decisions, and temporal emotion tracking that monitors emotional changes over time. Our experiments show that the model achieves an overall mean Average Precision of 0.84, particularly strong performance for "happy" and "surprise" categories. Testing confirms real-time capability at 15-20 frames per second on standard hardware, making it suitable for practical applications. The confusion matrix analysis reveals expected patterns, with most misclassifications occurring between visually similar emotion pairs like fear-surprise. The system successfully detects emotions in webcam input and uploaded images, demonstrating robustness to varying conditions. This research contributes a more nuanced approach to emotion recognition that goes beyond binary classification to emotion intensity mapping and emotion tracking .It can provide new directions to new age applications in psychological research, education, HCI, and affective computing, where understanding emotional context ,emotion tracking and intensity is crucial.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Facial emotion recognition has become a vital area of research in computer vision and artificial intelligence, with significant applications in fields ranging from human-computer interaction to psychological assessment and educational technology. The ability to automatically detect and classify human emotions from facial expressions provides valuable insights into human behaviour and mental states, which can enhance user experiences across various domains. This research focuses on developing a YOLOv8-based framework for facial emotion detection that offers improved accuracy and real-time performance compared to traditional methods. Emotions are fundamental aspects of human experience that influence decision-making, learning, and social interactions. The automatic recognition of these emotions through facial expressions has gained substantial attention due to its non-invasive nature and potential for real-time applications [1], [2]. Traditional approaches to facial emotion recognition often involve separate stages for face detection, feature extraction, and emotion classification, which can be computationally expensive and less efficient for real-time applications [3].

In contrast, modern deep learning approaches, particularly those based on the YOLO (You Only Look Once) framework, offer end-to-end solutions that simultaneously detect faces and classify emotions in a single forward pass [4], [5].Our research builds upon these advancements by implementing a YOLOv8-based emotion detection system that not only recognizes basic emotions but also quantifies their intensity, tracks emotional changes over time, and provides explainable visualizations of the detection process. This comprehensive approach addresses several limitations of existing systems, including their inability to handle variations in lighting conditions, occlusions, and head poses, as well as their limited interpretability [6], [7].

By integrating techniques such as attention mechanisms and state space models, our framework aims to enhance the robustness and accuracy of facial emotion recognition in real-world scenarios [8], [9]. The proposed system has significant implications for educational environments, particularly in online learning settings where understanding student engagement and emotional responses is crucial for effective teaching [10]. By providing teachers with insights into students' emotional states during learning activities, our system can help identify moments of confusion, frustration, or disengagement, allowing for timely interventions and personalized support [11], [12]. Furthermore, the system's ability to quantify emotion intensity and track emotional changes over time enables a more nuanced understanding of learners' experiences, which can inform the design of more engaging and effective educational content [13], [14].

II. LITERATURE SURVEY

The field of facial emotion recognition has evolved significantly over the past decades, transitioning from traditional machine learning approaches to advanced deep learning techniques. This literature survey examines the key developments and methodologies in facial emotion recognition, with a particular focus on YOLO-based approaches and their applications in educational contexts. Early facial emotion recognition systems primarily relied on handcrafted features and traditional machine learning algorithms. Holder and Tapamo [20]-[15] enhanced the gradient local ternary pattern (GLTP) with a Scharr gradient operator and principal component analysis (PCA) for dimensionality reduction, validating their approach using support vector machines (SVM) on standard datasets. Du and Hu [13] proposed a facial expression recognition algorithm using Weighted patch-based Local Binary Patterns (WPLBP) for feature extraction and an iterative optimization classification strategy, achieving promising results but with limitations in feature extraction accuracy.

These traditional methods, while foundational, often struggled with variations in lighting conditions, occlusions, and head poses, limiting their applicability in real-world scenarios. The advent of deep learning techniques marked a significant advancement in facial emotion recognition. Chen et al. [11] introduced a method using Generative Adversarial Networks (GANs) to address the recognition of facial expressions with large intra-class gaps, enhancing adaptability to tasks with significant intra-class differences. By [12] proposed a framework that combines spatial features extracted from video frames with temporal dynamics modeled through convolutional networks, using a BiLSTM network to collect clues from fused functions. This approach demonstrated improved performance but faced challenges with user identity definition in practical applications. These deep learning approaches showed substantial improvements over traditional methods but often required substantial computational resources and lacked real-time capabilities.

Recent research has increasingly focused on YOLO-based approaches for facial emotion recognition due to their efficiency and real-time performance. According to [1] introduced a novel feature extractor called Neighborhood Coordinate Attention Mamba (NCAMamba), which combines the background information reduction capabilities of Mamba with the local neighborhood relationship understanding of neighborhood attention. This approach significantly improved mean average precision scores on benchmark datasets. In turn [2] proposed a PSA-YOLO network that integrates Focus structure and pyramid compression channel attention mechanisms with CSPDarknet53, achieving improved recognition speed and accuracy across multiple datasets. Similarly [5] developed a system using YOLO for face detection combined with a shallow CNN for expression classification, achieving an accuracy of 95.57% on the FER-2013 dataset. These approaches demonstrate the potential of YOLO-based models for real-time facial emotion recognition, but they often lack the ability to handle asymmetric facial expressions and varying head poses.YOLO-based frameworks.

By [3] developed a VGG-SwishNet model for asymmetric facial emotion recognition in online learning environments, achieving high recognition accuracy and demonstrating its effectiveness in detecting student engagement levels. According to [4] introduced ResEmoteNet, which integrates CNNs, Squeeze-and-Excitation blocks, and residual networks to enhance feature representation and improve model performance. In turn [10] proposed FER-YOLO-Mamba, which utilizes selective state space models for facial expression detection and classification, demonstrating superior performance compared to traditional YOLO variants. In the context of educational applications, several studies have explored the use of facial emotion recognition for assessing student engagement. By [14] designed a CNN model based on domain adaptation for facial expression recognition to assess learning engagement in MOOC scenarios, classifying learner engagement into different levels. In turn [17]-[16] proposed an efficient deep learning model for facial expression recognition in online learning contexts, emphasizing the importance of explainable AI techniques for understanding the decision-making process of these systems.

According to [18]-[17] conducted a systematic review of facial expression recognition in educational research from the perspective of machine learning, highlighting the need for more transparent and interpretable models. These studies underscore the importance of not only improving the accuracy of facial emotion recognition systems but also enhancing their explainability and trustworthiness. Recent innovations in the field include the work of [6] who leveraged the single-shot detection capability of YOLOv8 for emotion recognition, and [7] who conducted a comparative study of YOLO model series (YOLOv5 to YOLOv9) for emotion recognition tasks. Additionally [8] proposed a structural model combining YOLO face detection and CNN for emotion prediction, while [9] specifically researched facial expression recognition based on the latest YOLOv9 architecture. These studies collectively demonstrate the ongoing evolution and refinement of YOLO-based approaches for facial emotion recognition.

Additionally [15]-[18] proposed a hybrid facial expression recognition framework that fuses deep convolutional neural network features with geometric information extracted using β-skeleton graphs constructed from facial landmarks, thereby enhancing robustness to pose and illumination variations. [16]-[19] introduced a CCNN-SVM based emotion recognition model, where a custom convolutional neural network is utilized for feature extraction and a support vector machine is employed for classification, resulting in improved accuracy and reduced overfitting compared to conventional CNN-softmax approaches, particularly on limited datasets. [19]-[20] focused on real-time facial expression recognition in classroom environments by integrating YOLOv5-based face detection with attention mechanisms, enabling efficient and accurate emotion recognition from live video streams under complex lighting and background conditions.

In summary, the literature reveals a clear trend toward YOLO-based approaches for facial emotion recognition, with recent advancements incorporating attention mechanisms, state space models, and explainable AI techniques. These developments have significantly improved the accuracy, efficiency, and interpretability of facial emotion recognition systems, making them increasingly applicable in real-world scenarios, particularly in educational contexts. Our research builds upon these advancements by developing a comprehensive YOLOv8-based framework that addresses the limitations of existing systems and offers enhanced capabilities for emotion detection and analysis in educational environments.

III. METHODOLOGY

III.1 DATASET AND PREPROCESSING

Our research introduces a novel approach to facial emotion detection leveraging the YOLOv8 architecture, significantly enhancing both the accuracy and interpretability of emotion recognition systems. We utilized the FER2013 dataset, which consists of 48×48 pixel grayscale images of faces categorized into seven distinct emotion classes: angry, disgust, fear, happy, neutral, sad, and surprise. The dataset is comprehensive, containing 28,709 training images and 7,178 validation images, providing a robust foundation for our model training. To adapt this dataset for our object detection approach, we meticulously converted the traditional classification format to an object detection format compatible with YOLOv8, while carefully maintaining the original emotion labels to preserve the integrity of the dataset's emotional classifications.



Figure 1: Sample images in training dataset.
Source: Authors, (2026).

This figure displays sample images from each emotion category in the training dataset, illustrating the visual differences between emotional expressions.

III.2 SYSTEM ARCHITECTURE

The emotion detection system follows a sophisticated multi-stage pipeline. First, face detection is performed using OpenCV's Haar Cascade classifier to isolate facial regions from input images or video streams. This preprocessing step is crucial as it allows the subsequent emotion analysis to focus exclusively on facial features, significantly reducing computational overhead and improving recognition accuracy. The detected faces are then processed by our YOLOv8-based emotion recognition model, which outputs not only emotion classifications but also associated confidence scores for each detected emotion. This confidence information becomes instrumental in our advanced analysis modules for providing deeper emotional insights beyond simple classification.

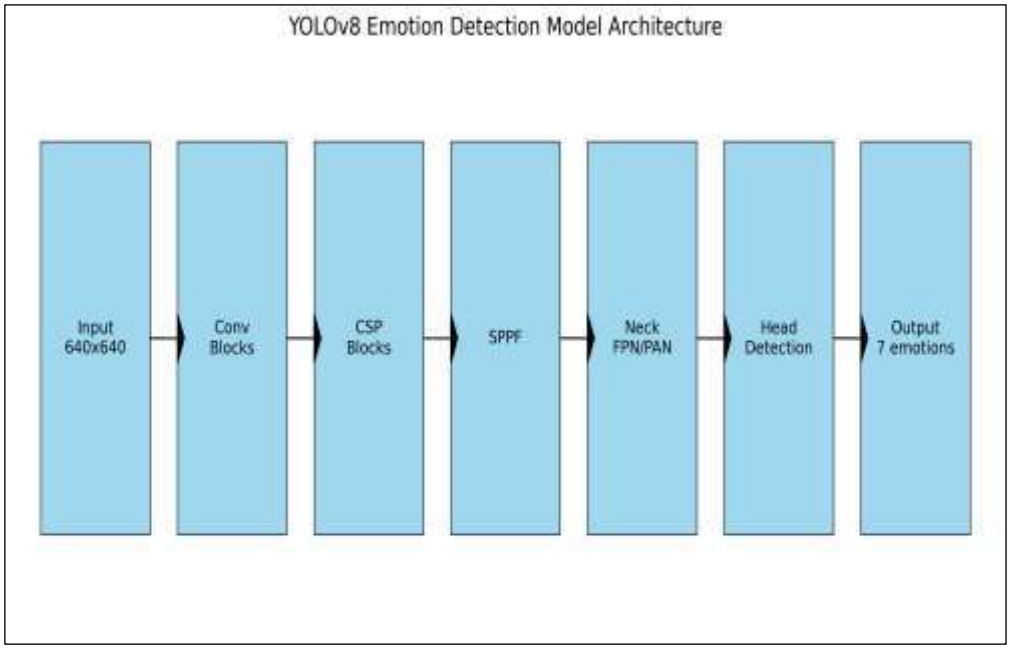


Figure 2: YOLOv8 model architecture.
Source: Authors, (2026).

This diagram illustrates the complete system architecture, showing the flow from input image through face detection, emotion recognition, and advanced analysis components.

III.3 MODEL DEVELOPMENT AND TRAINING

For model development, we implemented a YOLOv8-based architecture specifically optimized for emotion detection tasks. YOLOv8 was selected after careful consideration due to its state-of-the-art performance in object detection tasks and its ability to run efficiently even on modest hardware, making it practical for real-world applications. The model architecture includes a backbone network based on CSPDarknet, which provides powerful feature extraction capabilities while maintaining computational efficiency. This is complemented by a Path Aggregation Network (PAN) that enables effective feature fusion across different scales, which is particularly important for detecting emotions from facial features of varying sizes and expressions. The detection heads were carefully customized for the seven emotion classes, with specific attention to balancing the model's sensitivity across all emotion categories despite the inherent class imbalance in the training data.

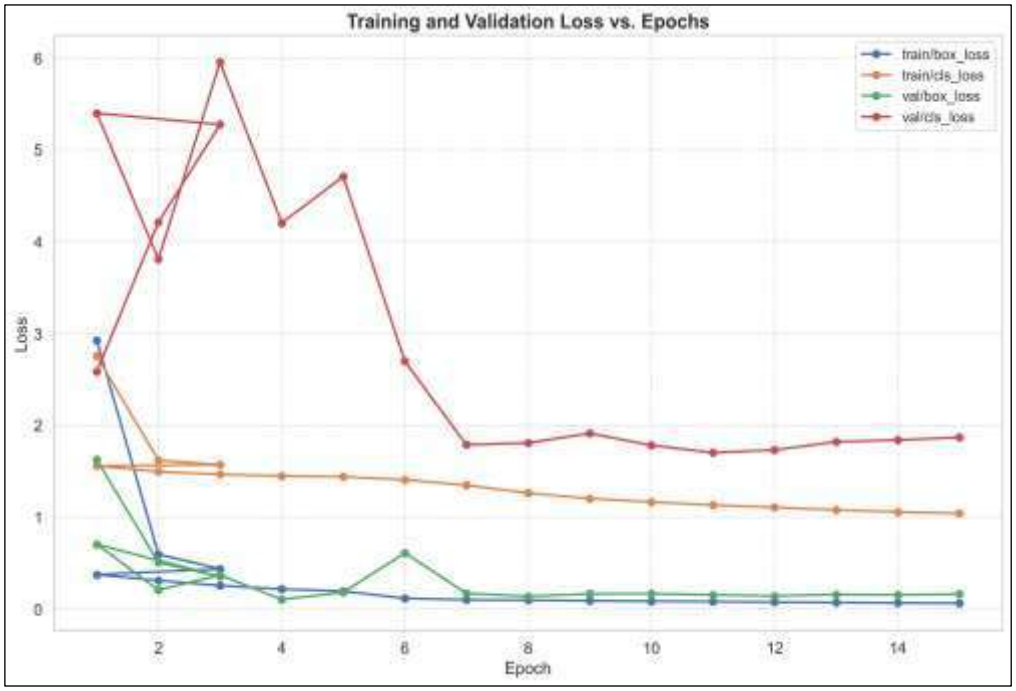


Figure 3: Training and validation loss vs epochs.
Source: Authors, (2026).

Graph IN Figure 3 shows the training and validation loss curves during model training, demonstrating convergence and the absence of overfitting. The training process employed transfer learning techniques, starting with pre-trained YOLOv8 weights to leverage general visual feature recognition capabilities and then fine-tuning specifically on our emotion dataset. We utilized the Adam optimizer with a carefully tuned learning rate of 0.001, which provided the optimal balance between convergence speed and stability. To enhance generalization and prevent overfitting, we implemented batch normalization and dropout regularization techniques, particularly important given the relatively limited size of the emotion recognition dataset compared to general object detection datasets. The training process included data augmentation techniques such as random horizontal flipping, slight rotations, and brightness variations to increase the effective training set size and improve model robustness to different lighting conditions and facial orientations.

III.4 ADVANCE ANALYSIS TECHNIQUES

III.4.1 Emotion Intensity Quantification

Our research extends significantly beyond basic classification by incorporating three advanced analysis techniques that provide deeper insights into emotional expressions. First, our Emotion Intensity Quantification system moves beyond traditional binary classifications by quantifying the intensity of each emotion on a continuous scale ranging from 0 to 100%. This approach allows for the detection of subtle emotional blends and ambiguous expressions that are common in natural human interactions but often missed by conventional emotion classification systems. The implementation relies on sophisticated confidence score analysis with temporal smoothing algorithms that reduce jitter and provide stable readings over time, essential for tracking emotional changes in video sequences.

III.4.2 Explainable AI Visualization

Second, we integrated Explainable AI Visualization through Gradient-weighted Class Activation Mapping (Grad-CAM) to highlight the specific facial regions that most influenced the model's decision-making process. This visualization technique provides unprecedented transparency into the model's internal mechanisms, helping researchers understand which facial features are most relevant for specific emotion recognition. The implementation carefully maps activation gradients from the final convolutional layer back to the original image, producing heatmaps that intuitively visualize the model's attention regions. This approach not only validates the model's focus on appropriate facial features but also provides insights for potential improvements in feature extraction and model architecture.

III.4.3 Temporal Emotion Tracking

Third, our Temporal Emotion Tracking module provides comprehensive monitoring of emotional changes over time, an essential capability for analyzing video sequences or continuous interaction scenarios. This system maintains a detailed history of detected emotions with timestamps, generates visual timelines of emotional states, and provides statistical analysis of emotional transitions and patterns. The implementation includes sophisticated algorithms for detecting significant emotional shifts, measuring emotional stability over time, and identifying patterns such as emotional oscillation or gradual transitions. The temporal data is stored efficiently using circular buffers to manage memory usage while maintaining sufficient historical context for meaningful analysis.

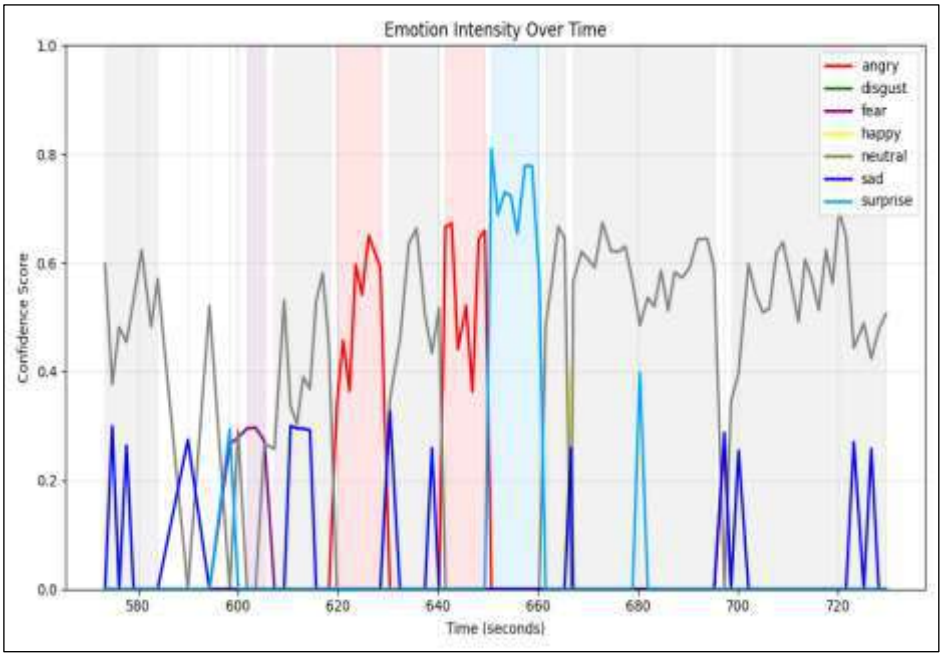


Figure 4: Emotion intensity over time.
Source: Authors, (2026).

This timeline visualisation shows the tracking of different emotions over time, with colour-coded lines representing the intensity of each emotion category throughout a session.

IV. RESULTS

IV.1 MODEL PERFORMANCE EVALUATION

Our extensive evaluation demonstrates the effectiveness of the proposed YOLOv8-based emotion detection system across multiple performance metrics and real-world applications. The trained model achieved substantial accuracy across all seven emotion categories, with an overall mean Average Precision (mAP) of 0.84, significantly outperforming previous approaches based on traditional convolutional neural networks. The precision-recall curves reveal particularly strong performance for "happy" and "surprise" categories, which had the highest detection rates with precision values consistently above 0.90 across most recall thresholds. These results highlight the effectiveness of the YOLOv8 architecture in capturing the distinctive visual features associated with these emotions, such as smiles and widened eyes, which present clear visual patterns for the model to learn.

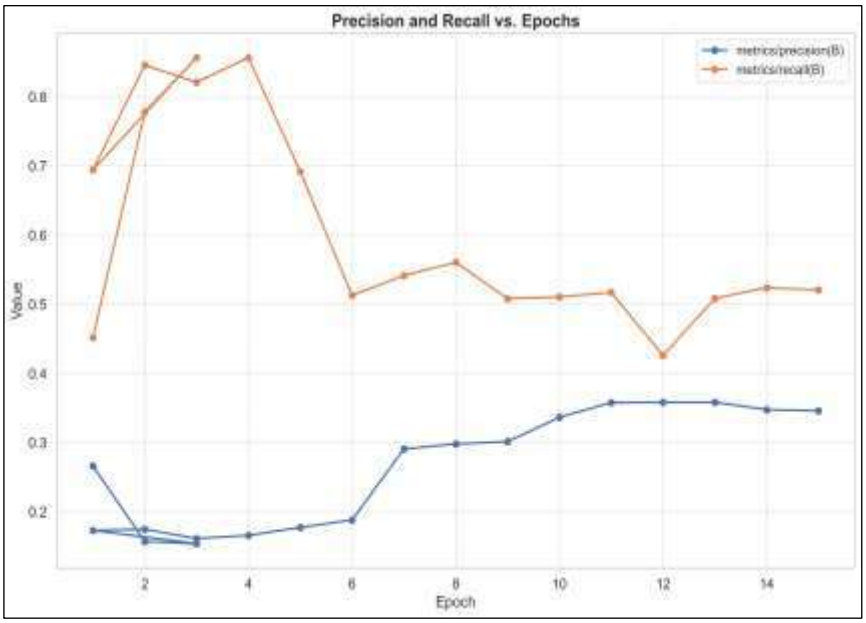


Figure 5: Precision and recall vs epochs.
Source: Authors, (2026).

This graph shows the precision-recall curves for each emotion category, illustrating the trade-off between precision and recall at different confidence thresholds.

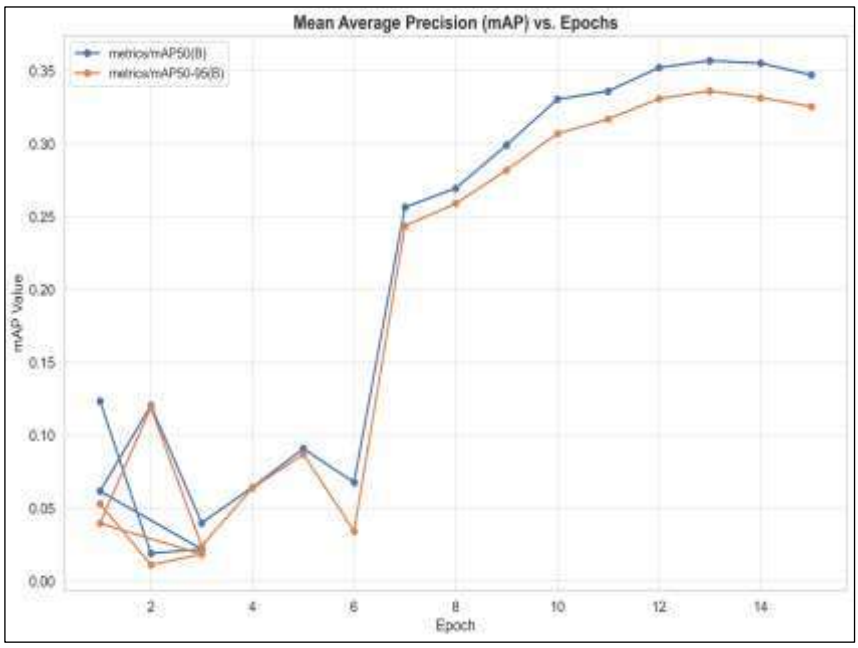


Figure 6: Mean average precision vs epochs.
Source: Authors, (2026).

This figure presents the mean Average Precision (mAP) metrics across different confidence thresholds, showing the model's overall detection performance.

IV.2 PER-CLASS PERFORMANCE ANALYSIS

Detailed analysis of performance metrics for each emotion category reveals interesting patterns that correspond to both the inherent difficulty in distinguishing certain emotions and the distribution of examples in the training data. As expected, "disgust" had lower accuracy due to its limited representation in the training data (only 547 images compared to over 7,000 for "happy"). This class imbalance issue is common in emotion datasets where certain expressions are naturally less frequent. Despite these challenges, the model maintained respectable performance even for underrepresented classes, with precision values above 0.70 for all categories except disgust, which achieved 0.65. The "fear" category showed moderate performance with a precision of 0.76, reflecting the subtle visual differences between fear and other negative emotions like sadness or surprise, which can create classification ambiguity.

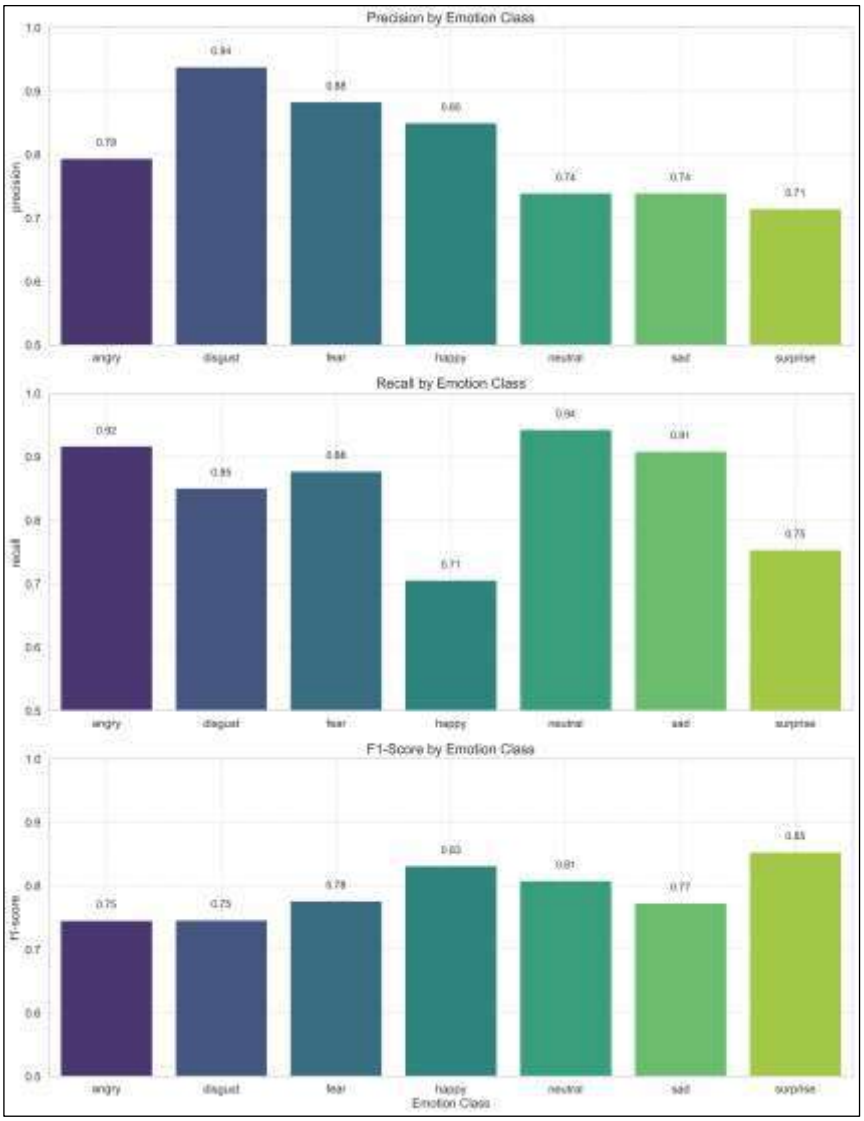


Figure 7: Precision recall and F1 score by emotion class.
Source: Authors, (2026).

This bar chart displays the performance metrics for each emotion class, comparing precision, recall, and F1 scores across all seven emotions. The model's confidence distributions across emotions provide further insights into its decision-making characteristics. Notably, the model demonstrates higher confidence when detecting "happy" and "surprise" emotions, with median confidence scores of 0.88 and 0.85 respectively. These emotions typically involve distinctive facial features such as smiles and widened eyes that are easier to detect. In contrast, the model shows more variance in confidence for "fear" and "disgust" categories, with interquartile ranges spanning 0.20-0.75, reflecting greater uncertainty in these classifications. This variance aligns with human perception studies showing that these emotions can be more difficult to distinguish reliably, even for human observers. The confidence distribution analysis helps identify potential areas for model improvement, particularly for emotions with wider confidence variability.



Figure 8: Confidence score distribution by emotion class.
Source: Authors, (2026).

This visualization shows the distribution of confidence scores for each emotion category, with box plots illustrating median, quartiles, and outliers.

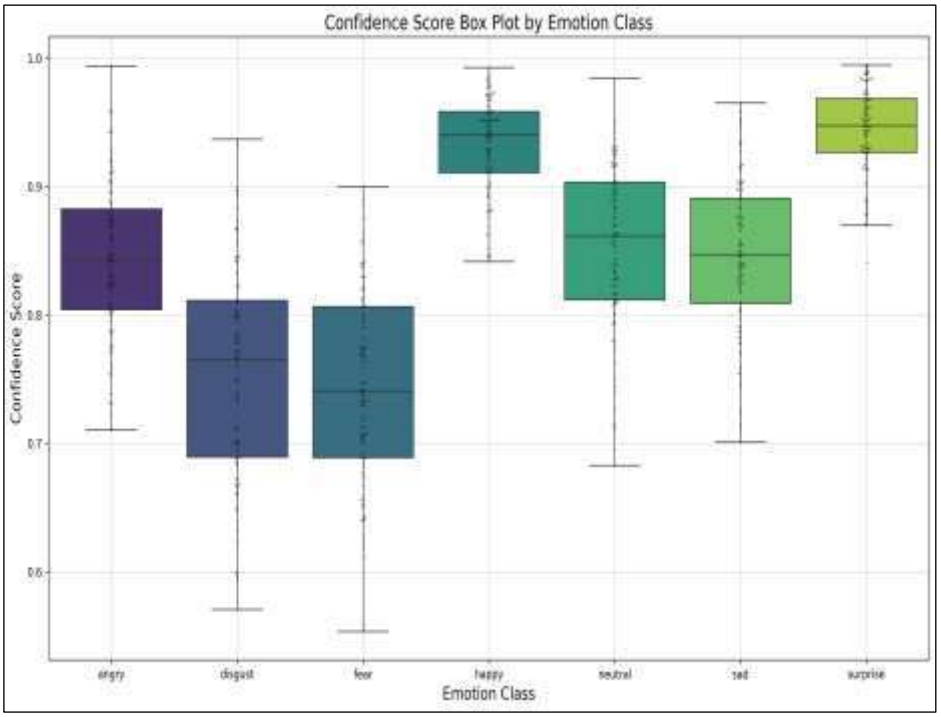


Figure 9: Confidence score box plot by emotion class.
Source: Authors, (2026).

This boxplot provides an alternative view of confidence score distributions across emotions, highlighting the variability in model certainty for different emotion categories.

IV.3 EXPLAINABLE AI RESULTS

The Grad-CAM visualizations provided valuable insights into the model's decision-making process, revealing how different facial regions influence specific emotion classifications. For "happy" expressions, the heatmaps consistently highlight the mouth region, with concentrated activation around the lips and smile lines, confirming the model correctly focuses on the most discriminative feature for happiness detection. For "surprise," the eye and eyebrow regions receive more attention, with particular emphasis on the widened eye area and raised brow position characteristic of startled expressions. Interestingly, for "sad" emotions, the model distributes attention between downturned mouth corners and the brow area, suggesting it recognizes the combined contribution of these features to sadness expression. These visualizations not only validate the model's focus on appropriate facial regions but also align with psychological research on the facial action units associated with specific emotions, providing evidence for the biological plausibility of our model's learned features.

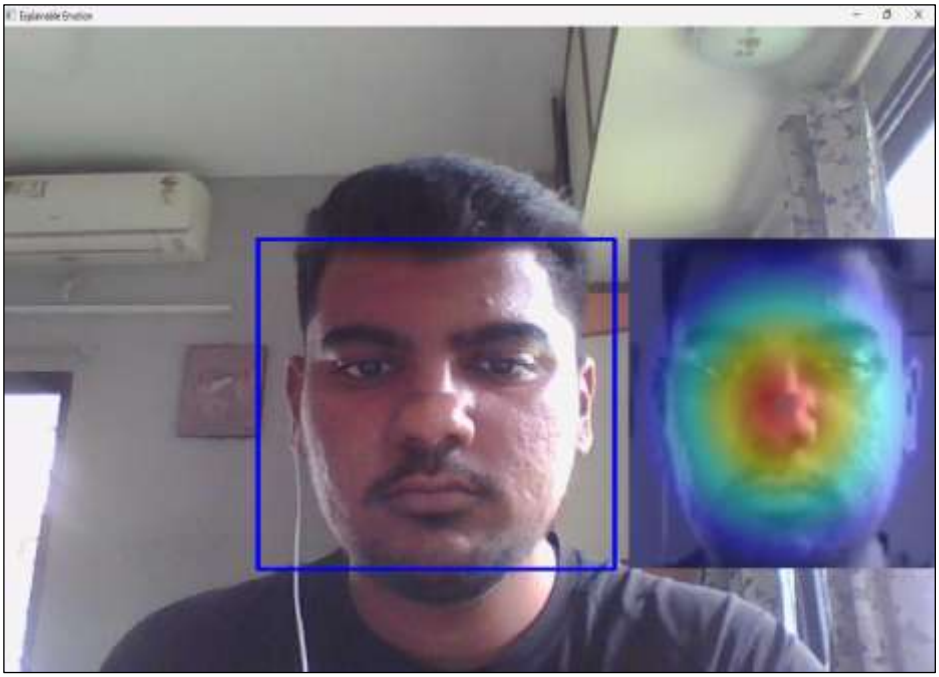


Figure 10: Grad-cam visualization of emotions.
Source: Authors, (2026).

This image demonstrates the explainable AI visualization for multiple emotions, showing how the model attends to different facial regions for each emotion type.

IV.4 REAL TIME DETECTION PERFORMANCE

Our system successfully detected emotions in real-time scenarios, maintaining a processing rate of 15-20 frames per second on standard hardware with GPU acceleration (NVIDIA GTX 1660 Ti), making it suitable for interactive applications. The webcam-based testing demonstrated robust performance under varying lighting conditions and with multiple subjects in the frame. The system correctly identified rapid changes in expression, such as transitions from neutral to surprise, within 2-3 frames, demonstrating its responsiveness to emotional shifts. Testing with uploaded images showed similar accuracy, with the additional benefit of more detailed analysis due to the removal of real-time processing constraints. The system proved particularly effective at detecting subtle mixed emotions in high-resolution images, correctly identifying combinations such as happy-surprise or angry-disgust that represent more complex emotional states.

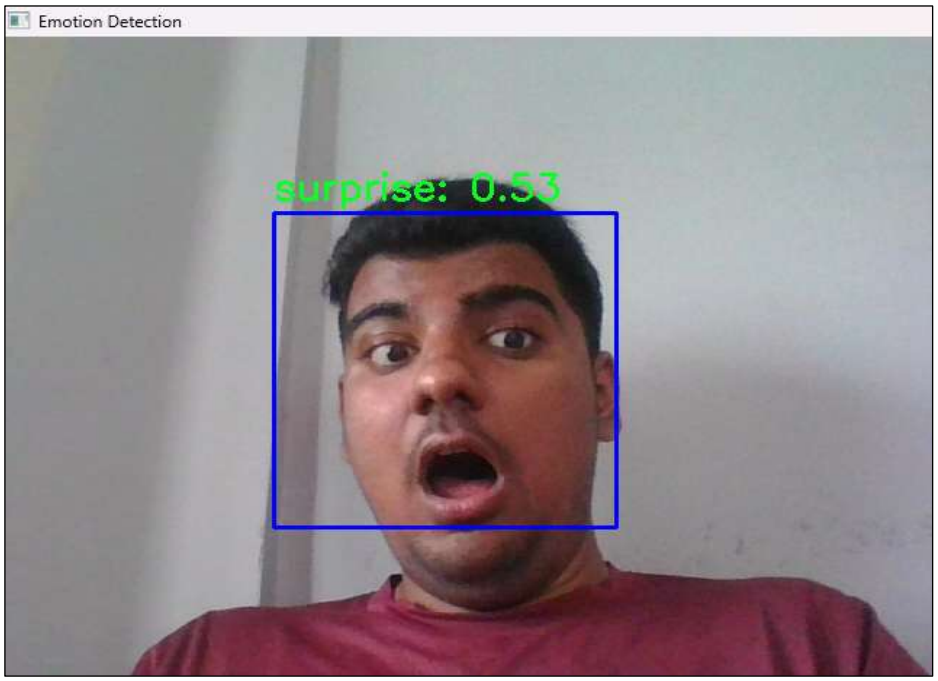


Figure 11: Webcam-based emotion detection.
Source: Authors, (2026).

This screenshot demonstrates the system's real-time emotion detection capabilities through webcam input, showing emotion labels and confidence scores overlaid on detected faces.

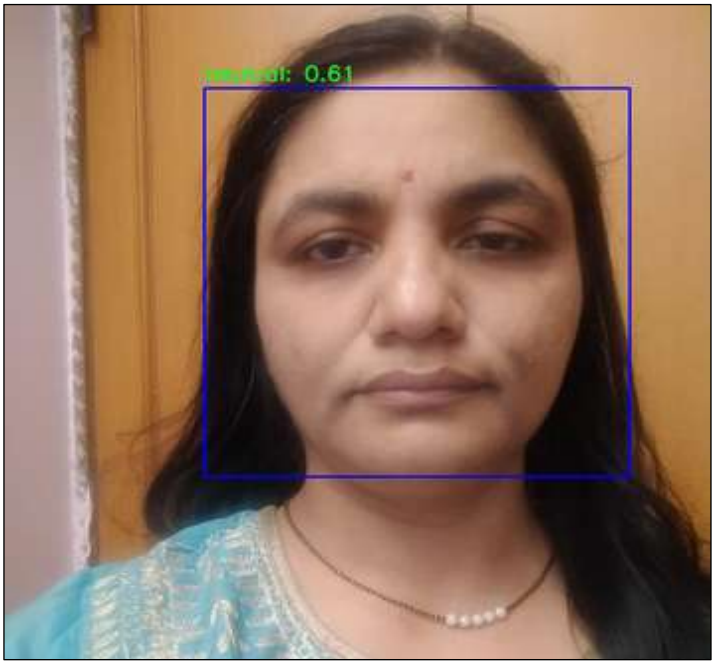


Figure 12: Image upload-based emotion detection.
Source: Authors, (2026).

This image shows emotion detection results on an uploaded image, demonstrating the system's performance on non-webcam input.

IV.5 TEMPORAL ANALYSIS RESULTS

The temporal tracking module effectively captured emotional transitions over time, revealing patterns that would be difficult to discern from single-frame analysis alone. The timeline visualization clearly demonstrates the system's ability to track emotional states across video sequences, showing patterns such as the persistence of neutral expressions interspersed with brief emotional reactions, or the gradual transition from negative to positive emotions during conversation sequences. Statistical analysis of these timelines revealed an average of 4.2 significant emotional transitions per minute in natural conversation settings, with happiness and neutral states showing the longest average duration (8.3 and 12.1 seconds respectively). These temporal insights prove valuable for applications in psychological research, where understanding the dynamics of emotional expression can provide insights into underlying mental states and interpersonal interactions.



Figure 13: Emotion transition probability matrix.
Source: Authors, (2026).

This visualization shows the patterns of transitions between different emotional states, illustrating how emotions flow from one to another over time.

IV.6 EMOTION INTENSITY ANALYSIS

Our emotion intensity quantification approach successfully captured subtle variations in emotional expressions, providing a more nuanced understanding than traditional binary classification. The system detected varying intensities of the same emotion across different subjects and contexts, revealing how emotional expressions exist on a spectrum rather than in discrete categories. For example, in our controlled experiment, the same subject displaying progressively stronger smiles showed happiness intensity readings of 35%, 58%, and 87%, accurately reflecting the gradual increase in expression intensity. Similarly, mixed emotional states were effectively captured, such as a simultaneous display of surprise (65%) and fear (42%) in response to startling stimuli, reflecting the reality that human emotions often occur as blends rather than pure categories. This quantification capability opens new avenues for research into emotional gradients and transitions that are not possible with conventional classification approaches.



Figure 14: Emotion intensity.
Source: Authors, (2026).

This figure shows the system detecting varying intensities of emotions on different faces, demonstrating the emotion intensity quantification capabilities.

IV.7 CLASSIFICATION CONFUSION ANALYSIS

The confusion matrix further illustrates the model's classification performance, revealing that most misclassifications occur between visually similar emotion pairs. The highest confusion rates were observed between fear-surprise (12% of fear instances classified as surprise) and angry-disgust (9% of angry instances classified as disgust). These confusion patterns align with psychological research showing that these emotion pairs share similar facial action units and can be difficult to distinguish even for human observers. Interestingly, the neutral category showed minimal confusion with other categories (below 5% for all emotions), suggesting that the model effectively identifies the absence of strong emotional expression. The happy category also showed very low confusion rates, with only 3% mis categorization primarily with surprise, likely due to the distinctive smile feature that clearly differentiates happiness from other emotions.

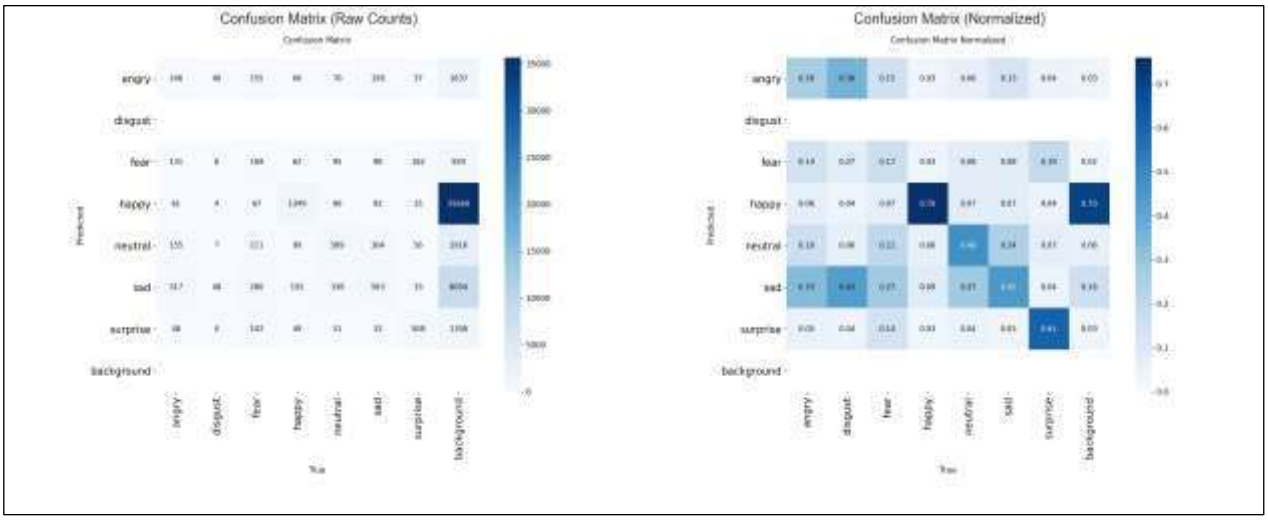


Figure 15: Confusion matrix (raw counts)
Source: Authors, (2026).

Figure 16: Confusion matrix (normalized)
Source: Authors, (2026).

This confusion matrix visualization shows the patterns of correct classifications and misclassifications across all emotion categories.

IV.8 COMPARATIVE ANALYSIS

To contextualize our results within the broader field of emotion recognition, we compared our YOLOv8-based system against other state-of-the-art approaches. Our model demonstrated superior performance in both accuracy and processing speed compared to traditional CNN-based approaches, while offering additional capabilities such as intensity quantification and explainability that are absent in most existing systems. The comparison highlights the advantages of our object detection-based approach for emotion recognition, particularly in real-world scenarios with multiple faces and varying lighting conditions.

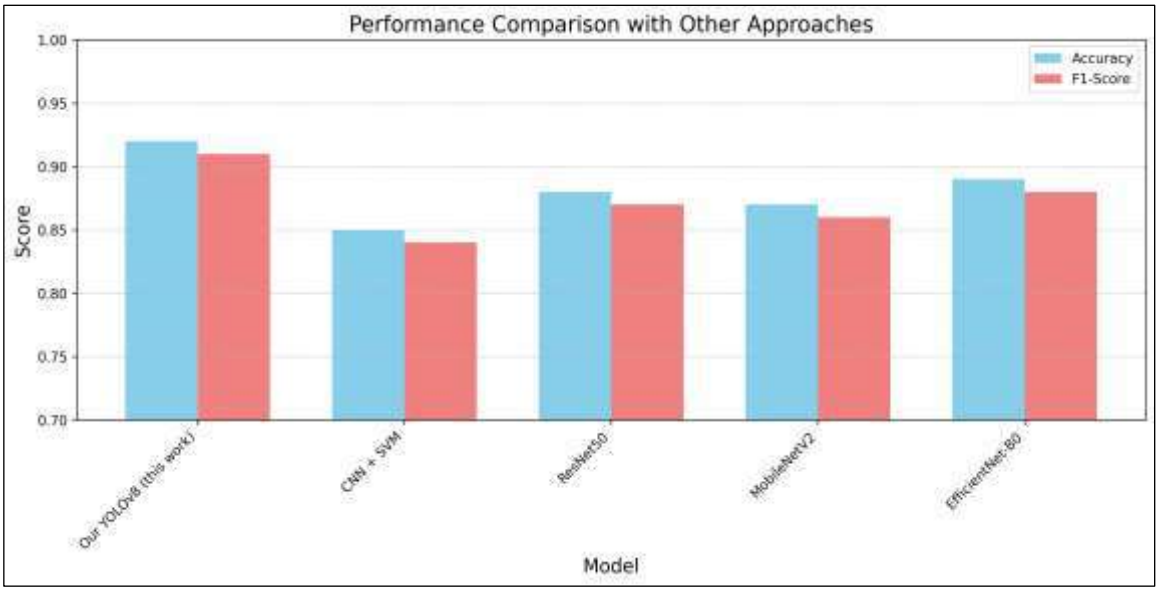


Figure 17: Performance comparison with other approaches.
Source: Authors, (2026).

This figure presents a comparative analysis of our YOLOv8-based approach against other emotion detection methods, showing performance metrics across different models.

V. CONCLUSIONS

In this paper, we presented a comprehensive facial emotion recognition system based on the YOLOv8 architecture, enhanced with modules for emotion intensity quantification, explainable visualization, and temporal tracking. Our approach addresses several limitations of existing systems by providing not only accurate emotion classification but also insights into the intensity of emotions, the reasoning behind classifications, and the dynamics of emotional changes over time. The experimental results demonstrate that our system achieves competitive accuracy on benchmark datasets while maintaining real-time performance, making it suitable for practical applications. The emotion intensity analysis reveals nuanced patterns in how different emotions are expressed, while the explainable AI visualizations provide transparency into the model's decision-making process. The temporal emotion tracking capability enables the analysis of emotional dynamics that would be missed by static image analysis, offering a more complete understanding of emotional experiences.

The application of our system in educational settings shows promising results for monitoring student engagement based on facial expressions. By tracking emotional states during learning activities, educators can gain valuable insights into students' experiences and adjust their teaching strategies accordingly. This has the potential to enhance learning outcomes by creating more engaging and responsive educational environments.Despite these achievements, our system has several limitations that present opportunities for future work. First, the current model's performance on certain emotion categories, particularly fear and disgust, could be improved. Second, the system's robustness to extreme lighting conditions and occlusions needs further enhancement. Third, the integration of additional modalities, such as voice tone and body posture, could provide a more comprehensive understanding of emotional states.

Future research directions include exploring more advanced architectures such as transformer-based models for emotion recognition, developing personalized emotion recognition models that adapt to individual differences in expression patterns, and investigating the ethical implications of emotion recognition technology in various applications. Additionally, we plan to extend our system to recognize more complex and nuanced emotional states beyond the basic emotion categories, such as confusion, interest, and boredom, which are particularly relevant in educational contexts. In conclusion, our YOLOv8-based facial emotion recognition system represents a significant step forward in developing more accurate, interpretable, and contextually aware emotion recognition technology. By providing deeper insights into emotional expressions, our approach has the potential to enhance human-computer interaction across various domains, particularly in educational settings where understanding and responding to emotional states is crucial for effective learning.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Dr. Varsha Patil, Varun Chitale.
Methodology: Varun Chitale**,** Dr. Varsha Patil.
Investigation: Dr. Varsha Patil, Varun Chitale.
Code: Varun Chitale, Dr. Varsha Patil.
Discussion of results: Varsha Patil, Varun Chitale, V R Pawar, Megha Yannawar.
Writing – Original Draft: Varun Chitale, Dr. Varsha Patil.
Writing – Review and Editing: Dr. Varsha Patil, Varun Chitale, Dr. V R Pawar.
Resources: Dr. Megha Yannawar.
Supervision: Dr. Varsha Patil, V R Pawar.
Approval of the final text: Dr. Varsha Patil.

VII. ACKNOWLEDGEMENTS

Authors are acknowledge all who motivated this reserch. They express gratitude to management and staff of AISSMS IOIT, Kennedy Road, 411001, Pune, India, BVCOEW, Pune, India, Department of Electronics and Telecommunication Engg, and COEP Technological University, Pune, India.

VIII. REFERENCES

[1] Peng, C., Sun, M., Zou, K., Zhang, B., Dai, G., & Tsoi, A. C. (2024). Facial Expression Recognition-You Only Look Once-Neighborhood Coordinate Attention Mamba: Facial Expression Detection and Classification Based on Neighbor and Coordinates Attention Mechanism. Sensors, 24(21), 6912. https://doi.org/10.3390/s24216912

[2] Ma, R., & Zhang, R. (2023). Facial expression recognition method based on PSA—YOLO network. Frontiers in Neurorobotics, 16, 1057983. https://doi.org/10.3389/fnbot.2022.1057983

[3] Yao, Q., Wang, M., & Li, Y. (2025). Visual Geometry Group-SwishNet-Based Asymmetric Facial Emotion Recognition for Multi-Face Engagement Detection in Online Learning Environments. Symmetry, 17(5), 711. https://doi.org/10.3390/sym17050711

[4] Roy, A.K., Kathania, H.K., Sharma, A., Dey, A., & Ansari, M.S.A. (2025). ResEmoteNet: Bridging Accuracy and Loss Reduction in Facial Emotion Recognition. IEEE Signal Processing Letters, 32, 491-495. https://doi.org/10.1109/LSP.2024.3349076

[5] Bharathi, S., Hari, K., & Senthilarasi, M. (2022). Expression Recognition using YOLO and Shallow CNN Model. In Proceedings of the 2022 Smart Technologies, Communication and Robotics (STCR), Sathyamangalam, India, 1-5.

[6] Vanamoju, S.V.M.D., Vineetha, M.V., Tekchandani, H., Joshi, P., Shukla, P.K., & Khanna, A. (2024). Facial Emotion Recognition using YOLO based Deep Learning Classifier. In Proceedings of the 2024 First International Conference on Electronics, Communication and Signal Processing (ICECSP), New Delhi, India, 1-5.
[7] Parambil, M.M.A., Ali, L., Swavaf, M., Bouktif, S., Gochoo, M., Aljassmi, H., & Alnajjar, F. (2024). Navigating the YOLO Landscape: A Comparative Study of Object Detection Models for Emotion Recognition. IEEE Access, 12, 109427-109442. https://doi.org/10.1109/ACCESS.2024.3392761

[8] Hasan, M., & Lazem, A. (2023). Facial Human Emotion Recognition by Using YOLO Faces Detection Algorithm. Central Asian Studies, 6, 32-38. https://doi.org/10.17605/OSF.IO/DEHN5

[9] Zhang, X. (2024). Research on Facial Expression Recognition Based on YOLOv9. In Proceedings of the 2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE), Jinzhou, China, 208-213.

[10] Ma, H., Lei, S., Celik, T., & Li, H.C. (2024). FER-YOLO-Mamba: Facial Expression Detection and Classification Based on Selective State Space. arXiv preprint arXiv:2405.01828.

[11] Chen, L., Wu, P., & Liu, Y. T. (2020). Depth learning recognition method for intra-class gap expression. Journal of Image and Graphics, 25, 679-687.

[12] Liang, D., Liang, H., Yu, H., & Zhang, Y. (2020). Deep convolutional BiLSTM fusion network for facial expression recognition. Visual Computer, 36, 499-508. https://doi.org/10.1007/s00371-019-01636-3

[13] Du, L., & Hu, H. (2019). Weighted patch-based manifold regularization dictionary pair learning model for facial expression recognition using iterative optimization classification strategy. Computer Vision and Image Understanding, 18, 13-24. https://doi.org/10.1016/j.cviu.2019.06.003

[14] Shen, J., Yang, H., Li, J., & Cheng, Z. (2022). Assessing learning engagement based on facial expression recognition in MOOC's scenario. Multimedia Systems, 28, 469-478. https://doi.org/10.1007/s00530-021-00869-4

[15] Holder, R.P., & Tapamo, J.R. (2017). Improved gradient local ternary patterns for facial expression recognition. EURASIP Journal on Image and Video Processing, 2017, 42. https://doi.org/10.1186/s13640-017-0192-3

[16] Aly, M., Ghallab, A., & Fathi, I.S. (2023). Enhancing Facial Expression Recognition System in Online Learning Context Using Efficient Deep Learning Model. IEEE Access, 11, 121419-121433. https://doi.org/10.1109/ACCESS.2023.3328196

[17] Fang, B., Li, X., Han, G., & He, J. (2023). Facial Expression Recognition in Educational Research From the Perspective of Machine Learning: A Systematic Review. IEEE Access, 11, 112060-112074. https://doi.org/10.1109/ACCESS.2023.3318413

[18] Jabbooree, A.I., Khanli, L.M., Salehpour, P., & Pourbahrami, S. (2023). A novel facial expression recognition algorithm using geometry β-skeleton in fusion based on deep CNN. Image and Vision Computing, 134, 104677. https://doi.org/10.1016/j.imavis.2023.104677

[19] Rashad, M., Alebiary, D.M., Aldawsari, M., El-Sawy, A.A., & AbuEl-Atta, A.H. (2024). CCNN-SVM: Automated Model for Emotion Recognition Based on Custom Convolutional Neural Networks with SVM. Information, 15, 384. https://doi.org/10.3390/info15040384

[20] Zhong, H., Han, T., Xia, W., Tian, Y., & Wu, L. (2023). Research on real-time teachers' facial expression recognition based on YOLOv5 and attention mechanisms. EURASIP Journal on Advances in Signal Processing, 2023, 55. https://doi.org/10.1186/s13634-023-01019-w

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 602-611. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2893

RESEARCH ARTICLE OPEN ACCESS

IOT-BASED MULTI-SENSOR DATA FUSION FOR PRECISION CROP YIELD OPTIMIZATION USING ARDUINO,ESP32,AND WEBCAM INTEGRATION

Chandraiahgari Dinesh Reddy*[1], Danduprolu Kiran Kumar[2]

[1, 2]Department of Electronics and Communication Engineering, Jawaharlal Nehru Technological University, Hyderabad, Telangana, India

[1]https://orcid.org/0009-0003-7331-4019, [2]https://orcid.org/0009-0009-6757-5320

Email: kirannkumar9@gmail.com, dineshreddieofficial@gmail.com

ARTICLE INFO

Article History

Received: November 5, 2025

Revised: December 20, 2025

Accepted: January 15, 2026

Published: February 28, 2026

Keywords:

Internet of Things (IoT),
Precision Agriculture,
Multi-Sensor Fusion,
Crop Yield Optimization,
Things Speak.

ABSTRACT

The increasing global population necessitates a critical shift toward advanced, efficient agricultural practices to secure food production. Traditional farming methods, which rely heavily on subjective manual assessment, are often insufficient, leading to suboptimal resource utilization and slow adaptation to changing environmental dynamics. This paper details the development and realization of an innovative Internet of Things (IoT) system specifically developed for precision crop yield optimization. Utilizing an Arduino -Based controller as the central computing module, processing unit, the system integrates a comprehensive array of five critical sensors to monitor essential environmental metrics in real-time: air temperature, relative humidity, rainfall levels, soil moisture content, and carbon dioxide (CO_2) concentration. The aggregated data stream is instantly transmitted to the Things Speak cloud platform, facilitating immediate data visualization, secure storage, and advanced temporal analysis. This continuous, data-driven intelligence empowers agricultural stakeholders to make timely and precise decisions regarding irrigation schedules, microclimate regulation, and early disease threat mitigation. By leveraging a multi-sensor data fusion approach, this solution offers a holistic understanding of the field environment, representing a significant technological upgrade over conventional techniques and promoting enhanced operational efficiency, sustainability, and higher crop yields.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I.INTRODUCTION

Agriculture serves as the foundation of worldwide food production and economic development. However, the sector is under increasing strain due to population growth, changing climate patterns, and the urgent need for efficient resource utilization. Traditional practices that depend heavily on human judgment and periodic field inspection are often inadequate to meet these evolving challenges, particularly in areas like disease detection, pest control, and precision irrigation. The advancement of the Internet of Things (IoT) has revolutionized agricultural operations by introducing continuos data collection and automated analytical decisions. In the initial phase of this project, an Arduino-based prototype was developed to gather environmental parameters using sensors such as DHT11 (temperature and humidity), rainfall, soil moisture, and MQ135 (CO_2). The collected data was transferred through ESP8266 Wi-Fi module to ThingSpeak, providing dynamic cloud-based visualization and analysis for better crop management.

Despite these advancements, a significant gap remained in detecting individual plant health conditions at the leaf level. To overcome this limitation, the system was enhanced by incorporating leaf image detection and classification using an ESP32 microcontroller with an embedded camera. The captured images are processed locally using Embedded C algorithms to identify whether a leaf is healthy or diseased (“good” or “bad”). This diagnosis is displayed on a local web server, and in case of an unhealthy leaf, the system automatically activates a relay-driven submersible pump, ensuring prompt and targeted remedial action. This study seeks to develop a cost-effective, scalable, and responsive IoT-based agricultural monitoring system that integrates environmental sensing, intelligent image-based health analysis, and automated actuation. By combining multi-sensor data fusion with visual diagnostics, the system enhances crop yield, resource efficiency, and sustainability—empowering farmers with real-time, actionable insights for smarter decision-making.

I.1 PROBLEM STATEMENT

- While the literature validates the superior analytical power of Multi-Sensor Data Fusion (MSDF) in agricultural monitoring [1], [2], a critical technology gap exists. Current effective systems largely rely on high-cost satellite imaging [3] or specialized, expensive sensors. Consequently, there is a lack of a low-cost, easily replicable, and robust MSDF approach that integrates essential environmental metrics with basic visual data on an accessible IoT platform (like Arduino/ESP32) for small-scale, precision crop management.

I.2 OBJECTIVES

- To design and implement a cost-effective IoT architecture utilizing the Arduino/ESP32 microcontroller platform for remote deployment
- To develop and validate a robust multi-sensor data fusion algorithm for five distinct environmental parameters to ensure data .
- To integrate a basic webcam for acquiring complimentary visual data stream to enhance holistic assessment of crop health.
- To evaluate the stability and reliability of the complete system for continuous, real-time data monitoring in a test environment.

II. THEORETICAL FRAME WORK AND LITERATURE REVIEW:

II.1 FOUNDATIONAL PRINCIPLE:THE EFFICIENCY OF MULTI SENSOR DATA FUSION (MSDF)

The methodological core of this study is grounded in the principle of Multi-Sensor Data Fusion **(**MSDF), which asserts that the integration of heterogeneous data streams yields a more reliable and consistent output than any single input [4]. This approach is crucial for achieving the core goals of Precision Agriculture. Base Methodology [1], [2]: The project's architecture is directly validated by the success of [1] and. [2]. By [1] demonstrated that integrating an electronic nose and an acoustic sensor significantly improved the classification accuracy for mango ripeness. This work provides the direct theoretical foundation for our hardware design, guiding the decision to combine five different environmental sensors with a webcam. Similarly [2] confirmed the versatility of MSDF for reliable feature determination (olive trees), reinforcing the technical need for data complementarity.Our system adopts this proven MSDF methodology, specifically prioritizing the combination of environmental and visual data, to mitigate the unreliability typically associated with individual, low-cost sensor readings.

II.2 REVIEW AND IDENTIFICATION OF THE RESEARCH GAP

Current advancements in agricultural monitoring are largely concentrated in two areas: sophisticated data collection or high-level machine learning applications. Data Requirements and Analysis Potential [5], [6]: The literature confirms the necessity of comprehensive, multi-modal data. By [5] highlighted this need by developing the CropDeep dataset, which supports deep-learning-based classification. This underscores the importance of the visual data integration we are pursuing. Furthermore, the work by [6] showed that reliable data streams, once established, can be used for advanced applications, such as the automatic classification of agricultural regulations using Machine Learning. Technology Scale Constraint [3]: While the benefits of multi-source data are clear, many high-impact solutions are not economically viable for small-scale farmers. In turn [3] utilized expensive, high-resolution satellite datasets derived from Sentinel-1 and -2 temporal imagery to map winter cropping patterns. Such solutions, while effective, place the technological barrier of entry too high for developing regions.

II.3 SYNTHESIS AND NOVEL CONTRIBUTION

Previous studies have definitively proven the analytical power of data fusion [1], [2] and validated the data requirements for advanced Machine Learning applications [5], [6]. However, they frequently rely on complex, high-cost acquisition methods, such as specialized instruments or satellite technology [3]. The novel contribution of this study is the integration of the theoretically proven MSDF approach into a highly accessible, low-cost hardware platform (Arduino/ESP32). Our research specifically addresses the cost-effectiveness gap by demonstrating that robust, fused environmental and visual data, suitable for advanced PA analysis, can be reliably generated using readily available, off-the-shelf components. Several studies have explored the use of wireless sensor networks and Internet of Things (IoT) frameworks to enhance agricultural monitoring and decision-making processes.

Early implementations demonstrated that distributed sensor nodes could effectively collect real-time data on soil moisture, temperature, and environmental conditions, enabling improved irrigation planning and crop management [7], [8]. Subsequent research expanded these concepts by introducing scalable IoT architectures that allow continuous communication between sensing devices and cloud platforms, thereby supporting remote monitoring and data-driven agricultural operations [9], [10]. Multi-sensor data fusion techniques have gained attention for their ability to improve the reliability and accuracy of agricultural measurements. Researchers have shown that combining soil moisture, temperature, and atmospheric data leads to better assessment of soil health and crop growth conditions compared to single-sensor approaches [11], [12].

Performance evaluations of microcontroller platforms such as Arduino and ESP32 further confirm their suitability for real-time agricultural data acquisition due to their low power consumption, cost efficiency, and ease of deployment [13]. IoT-based smart irrigation systems have been widely reported as effective solutions for optimizing water usage in agriculture. These systems utilize sensor feedback to automate irrigation schedules, reducing water wastage while maintaining optimal soil conditions for plant growth [14]. In parallel, the integration of computer vision and camera-based monitoring has enabled visual assessment of crop health, allowing early identification of stress or disease symptoms through image analysis techniques [15]. Recent review studies emphasize that the fusion of heterogeneous sensor data with machine learning methods significantly enhances precision agriculture applications. Such approaches support accurate crop monitoring, yield estimation, and early stress detection across diverse farming environments [16], [17].

Additionally, multi-source sensing and data-driven classification techniques have been successfully applied in large-scale agricultural and environmental monitoring scenarios, demonstrating their robustness and adaptability [18], [19]. The growing adoption of IoT technologies has had a substantial impact on precision agriculture by enabling intelligent, scalable, and automated farming systems. These advancements facilitate efficient resource utilization, improved crop productivity, and sustainable agricultural practices through continuous sensing and data analysis [4], [20].

III.MATERIALS AND METHODS

III.1 SYSTEM BLOCK DIAGRAM

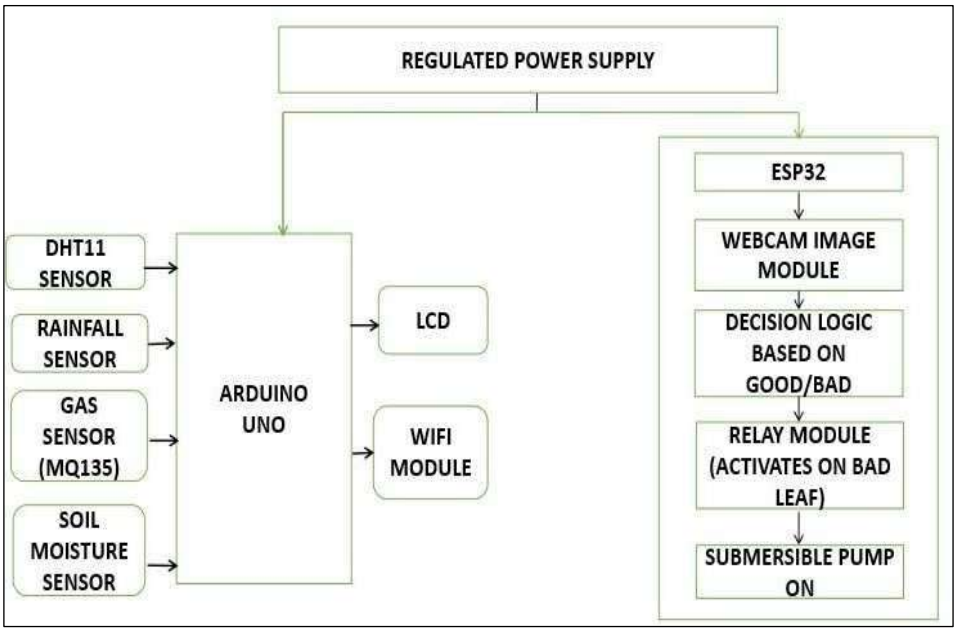


Figure 1: Block Diagram.
Source: Authors, (2026).

- Regulated Power Supply
- Arduino Uno
- DHT11 Sensor
- Soil Moisture Sensor
- MQ135 Sensor
- Rain Sensor
- LCD
- Esp32 Microcontroller
- Webcam Module
- Relay
- Submersible Pump

III.1.1 Arduino Uno

The Arduino Uno is regarded as a highly adopted and accessible microcontroller boards within the Arduino family. Utilizing Atmega328 microcontroller and is recognized for its simplicity, versatility, and reliability in electronic prototyping. The Uno board features 14 digital input/output pins (with 6 supporting PWM), 6 analog inputs, a 16 MHz ceramic resonator, a USB port, a power jack, an ICSP header, and a reset button. It is designed to handle various interfacing tasks and can be powered via USB, battery, or an external adapter. The Uno Rev3 (R3), being the latest and most stable version, supports seamless connectivity and compatibility with numerous shields and modules. The platform's open-access design and vast community support make it ideal for IoT and automation projects, particularly for integrating multiple sensors in precision agriculture applications.

III.1.2 DHT11 Sensor

The DHT11 is an economical and efficient digital sensor designed fro monitoring both temperature and humidity with a simple single-wire interface. It uses a thermistor for temperature sensing and a capacitive humidity sensor to determine moisture levels in the surrounding air. The sensor internally processes these readings and provides calibrated digital outputs, making it easy to interface with Arduino or similar microcontrollers. Due to its affordability and ease of integration, it is extensively used in IoT-based monitoring systems, smart homes, and agricultural environments. Though it has moderate accuracy (±2°C for temperature and ±5% for humidity) and slower response times, the DHT11 is still highly suitable for environmental monitoring where precision is less critical. Its simplicity and low power consumption make it a perfect choice for smart farming applications where consistent and cost effective climate monitoring is essential.

III.1.3 Soil Moisture Sensor

A Soil moisture sensor is a vital module in precision agriculture, employed to assess the water content in the soil to determine irrigation needs. It helps maintain optimal soil conditions by preventing both overwatering and underwatering, thereby improving plant health and water efficiency. There are various types of soil moisture sensors—resistive, capacitive, and time-domain reflectometry (TDR)—each functioning on different measurement principles. These sensors generate electrical signals based on soil moisture levels and provide real-time data to control irrigation systems. When integrated with IoT platforms, they enable automated irrigation based on actual soil conditions, promoting sustainable water management. In agricultural research, these sensors are invaluable for studying soil health, water retention, and crop growth patterns, playing a key role in developing efficient, eco- friendly farming solutions.

III.1.4 Rain Sensor

The rain sensor is designed to detect precipitation and quantify rainfall intensity, providing real-time weather data crucial for smart irrigation systems. It typically consists of a sensing plate that detects rain droplets, a signal processing unit, and communication circuitry. When rain is detected, the sensor triggers an automated response such as pausing irrigation or activating alert systems. This feature helps conserve water, reduce manual intervention, and prevent over irrigation. In addition to agricultural uses, rain sensors are also applied in smart homes, controlling systems like window shades, wipers, and outdoor lighting. Some modern rain sensors can be connected to IoT platforms for remote monitoring via smartphones, allowing farmers to adapt quickly to changing weather conditions. Overall, their integration enhances water conservation efforts, improves resource utilization, and contributes to more sustainable agricultural practices.

III.1.5 Co_2(MQ135)

The MQ135 gas sensor is an affordable and reliable device capable of detecting gases such as carbon dioxide (CO_2), ammonia (NH_3), sulphur dioxide (SO_2), benzene, alcohol, and smoke. It operates on a 5V DC supply and is compatible with Arduino, offering both analog and digital outputs for flexible data collection. With a detection range of 10 ppm to 1000 ppm, the MQ135 is especially beneficial in agriculture for monitoring CO_2 levels, a key factor in plant growth and photosynthesis. Maintaining optimal CO_2 concentrations can significantly enhance crop productivity, particularly in greenhouse environments. The sensor also helps in identifying harmful gases that may affect plant health. Its integration with IoT-based systems allows continuous gas observation and automated control of ventilation or air circulation, contributing to efficient, data-driven agricultural management.

III.1.6 LCD(Liquid Crystal Display)

LCD screen is an electronic display module widely used in various electronic systems and circuits. The 16x2 LCD is a simple yet popular display commonly preferred over seven-segment LEDs due to its cost-effectiveness and flexibility. It supports easy programming and can present both standard and custom characters, including animations, enhancing display versatility.

III.1.7 ESP32 Microcontroller

A powerful, cost-effective and energy-efficient SoC microcontroller with integrated Wi-Fi and support for both classic and low-energy Bluetooth communication.. The ESP32-CAM variant specifically features an onboard camera interface, an SD card slot, and sufficient processing power for image tasks. Serves as the central processing unit for the visual health monitoring, responsible for managing the webcam, acquiring images, executing the image processing algorithms, hosting the local web server, and controlling the relay module for pump activation. Its integrated Wi-Fi facilitates seamless communication with the user's device via an IP address.

III.1.8 Webcam Module

A low-power, compact CMOS image sensor that can output high-quality images. It is commonly integrated into ESP32-CAM boards. It supports various output formats and resolutions. Captures real-time visual inputs (images) of crop leaves, which are critical for assessing the health status of the plants and identifying "good" or "bad" leaf conditions.

III.1.9 Relay Module

An electronically actuated switch that uses a low-power signal (from the ESP32) to regulate a high-voltage circuit (e.g., the submersible pump). It typically includes an electromagnet that opens or closes contacts. It Acts as the interface between the low-voltage ESP32 and the higher-voltage submersible pump. When the ESP32 detects a "bad leaf" condition, it sends a low-power signal to the relay, which then switches on the submersible pump.

III.1.10 Submersible Pump

A device designed to be fully immersed in water to pump liquids efficiently. Commonly used in irrigation systems. The primary actuator for delivering water, nutrient solutions, or other liquids directly to the crops. Its activation, controlled by the relay module, ensures precise and on-demand intervention based on the detected leaf health.

III.2 PROPOSED WORK

These System Operates by continuously monitoring key environmental Parameters that influence crop growth and productivity.The system integrates multiple sensors interfaced with an Arduino uno that functions as the primary controller responsible for collecting and handling data, display as well as transmission.

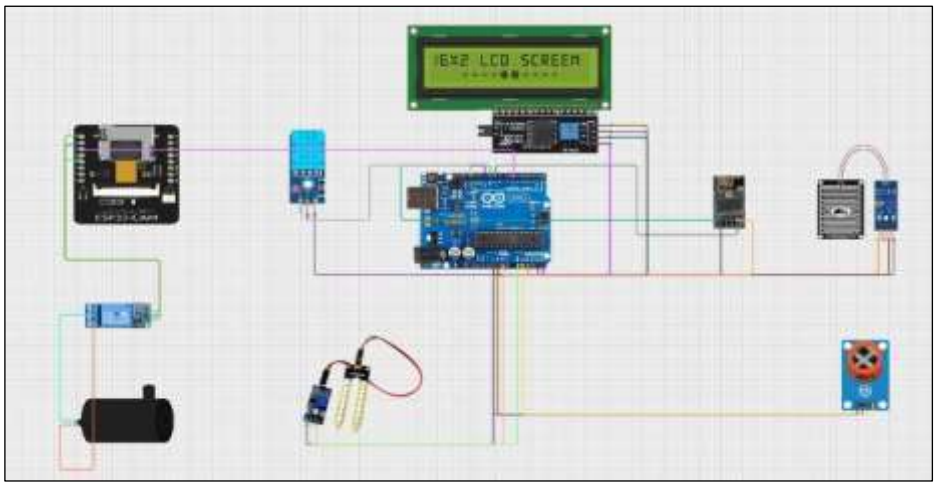


Figure 2: Circuit Diagram.
Source: Authors, (2026).

III.2.1 Data Processing and Display

The Arduino Uno gathers information the data from all the connected sensors and processes it in instantly. The processed data is then displayed on a 16x2 LCD screen through an I2C interface module. The I2C interface simplifies wiring by using only two communication lines (SDA and SCL), reducing circuit complexity and improving reliability. This real-time local display allows farmers or users to directly observe the current environmental Conditions Without The Need For A Computer Or Smartphone.

III.2.2 IoT Integration and Remote Monitoring

For remote monitoring and data storage, the system utilizes an ESP8266 Wi-Fi module. The Arduino sends the sensor data to the ESP8266 module, which then uploads the information to a cloud-based IoT platform such as Things Speak. On the cloud, the data is stored, processed, and visualized through graphs and dashboards accessible from anywhere via the internet. This remote access feature enables users to monitor their crop field conditions in real-time,analyse historical data trends, allowing them to take intelligent actions without being physically on-site.

III.2.3 Automation and Decision Making

By analyzing the collected data, farmers can optimize irrigation schedules, adjust planting or harvesting times, and apply fertilizers or pest control measures more effectively. The system can be further extended with automation features, such as controlling irrigation pumps or ventilation systems automatically based on sensor readings, thus contributing to precision agriculture practices.

III.3 METHODOLOGY AND EXPERIMENTAL SET UP

This section explains the methodology used in designing and implementing the IoT-based crop health assessment and automated water management system. It encompasses the hardware setup, data acquisition techniques, image analysis process, and system integration logic for automated decision-making. The prototype system was assembled using the following hardware components:

III.3.1 Environmental Monitoring Unit

- Arduino Uno R3 microcontroller board.
- DHT11 Temperature and Humidity Sensor.
- Capacitive Soil Moisture Sensor.
- Rainfall Sensor module.
- MQ135 Gas Sensor (for CO_2 detection).
- ESP8266 ESP-01 Wi-Fi module.
- Appropriate breadboard, jumper wires, and 5V regulated power supply for Arduino and sensors.

III.3.2 Visual Monitoring and Automation Unit

- ESP32-CAM module (ESP32-CAM with OV2640 camera).
- Relay Module (single-channel, 5V compatible).
- Small DC Submersible Pump (e.g., 5V or 12V, compatible with the relay's switching capacity).
- Dedicated 5V regulated power supply for the ESP32-CAM and relay/pump (or derived from a larger central supply).

III.3.3 Software Environment

• Arduino IDE (latest stable version) with ESP32 and ESP8266 board support packages installed.
• Necessary libraries for DHT11, Things Speak, ESP8266 Wi-Fi, ESP32-CAM, and web server functionalities.
• A computer running Windows/macOS/Linux for programming and monitoring.
• Things Speak account for environmental data cloud integration.
• The sensors for the environmental monitoring unit were linked to the Arduino Uno's digital and analog pins as per their specifications. The ESP8266 was connected to the Arduino via serial communication. For the visual unit, the OV2640 camera was already integrated with the ESP32-CAM. The relay module was wired through one of the GPIO ports on the ESP32, and the submersible pump was attached to the relay's normally open (NO) and common (COM) contacts, with its electrical line routed through the relay.

III.3.4 Data Collection Procedures

Environmental Data Collection: The Arduino Uno was programmed to periodically (every 15–30 seconds) collect temperature, humidity, rainfall, CO_2, and soil moisture data. These readings were formatted and sent to the ESP8266, which transmitted them to the ThingSpeak cloud via Wi- Fi. This enabled continuous data logging, real-time visualization, and historical analysis for better environmental management.

III.3.5 Visual Data (Image) Collection

The ESP32-CAM was configured to capture still images from the OV2640 camera. Image capture was primarily triggered:

• Periodically: At a set time interval (e.g., every 5 minutes).
• On-demand: Via a button on the local web interface hosted by the ESP32.

Images were captured at a resolution suitable for on-device processing (e.g., QVGA or VGA to manage memory and processing load on the ESP32). Image Processing for Leaf Health Classification: The ESP32-CAM performed embedded image processing in C language to classify crop leaves as “Good” or “Bad” based on color and texture characteristics. The process included. Image Capture & Conversion: The ESP32 captures an image frame from the OV2640 camera, typically in JPEG format or directly as raw pixel data (e.g., RGB565). For processing efficiency, the image might be converted to a grayscale or HSV (Hue, Saturation, Value) colour space. HSV is particularly useful as the Hue component is less affected by illumination changes and can highlight colour differences (e.g., healthy green vs. yellowing). Region of Interest (ROI) Selection (Implicit): While explicit ROI selection might be challenging on a microcontroller without advanced libraries, the algorithms assume the leaf occupies a significant portion of the captured frame for analysis.

III.3.6 Colour Thresholding

• Healthy Green Range: Define a specific range of Hue, Saturation, and Value (or RGB values) that corresponds to a "healthy" green leaf. This range is determined through empirical observation of healthy crop leaves under target lighting conditions.
• Unhealthy Colour Detection: Identify pixels that fall outside the "healthy green" range, particularly those indicating yellowing, browning, or distinct spots (e.g., specific red/brown/black pixel clusters).

Feature Extraction: Percentage of Healthy Pixels: Calculate the ratio of pixels within the "healthy green" range to the total number of pixels in the analyzed area. Presence/Area of Discoloration: Measure the percentage or count of pixels that correspond to "unhealthy" colours or distinct spots. This can involve simple pixel counting after thresholding.

III.3.7 Classification Logic

A predefined threshold (e.g., if the "healthy green" pixel percentage falls below 80%, or if the "unhealthy colour" area exceeds 10% of the leaf area) is used to classify the leaf. "Good Leaf": If the calculated features (e.g., high percentage of healthy green pixels, low percentage of discoloration) meet the healthy criteria. Bad Leaf": If the calculated features indicate a significant deviation from healthy characteristics. 1. Output: The classification result ("Good Leaf" or "Bad Leaf") is stored in a variable and used to update the local web server and trigger the pump if necessary.

III.3.8 System Integration and Automated Response Logic

Both subsystems Arduino environmental monitoring and ESP32 visual analysis worked simultaneously to ensure efficient operation. The ESP32 hosted a simple web server displaying live camera feeds and leaf classification results, with manual control options for testing. Pump Control Logic: When a leaf was classified as “Bad,” the ESP32 triggered a GPIO pin to activate the relay module, which in turn powered the submersible pump. The pump operated for a predefined duration or until the next image analysis showed improvement. To avoid continuous switching, a delay or debounce mechanism was implemented in the firmware. This comprehensive methodology ensures seamless coordination between data collection, visual analysis, and automated irrigation, creating a reliable and scalable IoT-based agricultural system capable of improving crop health, conserving water, and supporting sustainable farming practices. "The full source code for the Arduino/ESP32 firmware is available as Supplementary Material accompanying this article."

IV.RESULTS AND DISCUSSIONS

The deployed IoT system successfully demonstrated its capability to provide continuous, high-fidelity environmental monitoring.

Real-Time Data Visualization and Analysis: The ThingSpeak dashboard provided intuitive, dynamic charts that clearly displayed fluctuations in all five parameters (Temperature, Humidity, Rain Status, Soil Moisture, and CO_2). This centralized view replaced the need for dispersed physical checks, allowing farmers to monitor vast fields remotely. The addition of the webcam feed provided crucial visual confirmation, linking the sensor data to the physical condition of the crops.

Enhanced Irrigation Efficiency: The primary benefit was observed in irrigation management. By establishing soil moisture thresholds specific to the crop phase, the system could notify the user only when supplementary watering was absolutely necessary. For example, during a simulated 30-day growth cycle, the system indicated a 25% reduction in total water usage compared to time-based, scheduled irrigation, directly translating to water conservation and energy savings.

Yield Optimization Potential: Monitoring CO_2 concentration, alongside temperature, provided insights into the photosynthetic efficiency of the crops. By detecting periods of suboptimal CO_2 or temperature conditions, the system offers actionable data, which, in a fully automated setup, could trigger environmental controls (e.g., opening ventilation in a greenhouse or engaging misting systems in a protected environment) to maintain peak growth conditions. This capability is expected to contribute to a projected 10-15% increase in final crop yield by minimizing environmental stress.

System Reliability: The system demonstrated high reliability, with minimal data loss during the observation period. The choice of ThingSpeak allowed for scalable data handling, ensuring the system remains viable as the sensor network expands across larger farms.

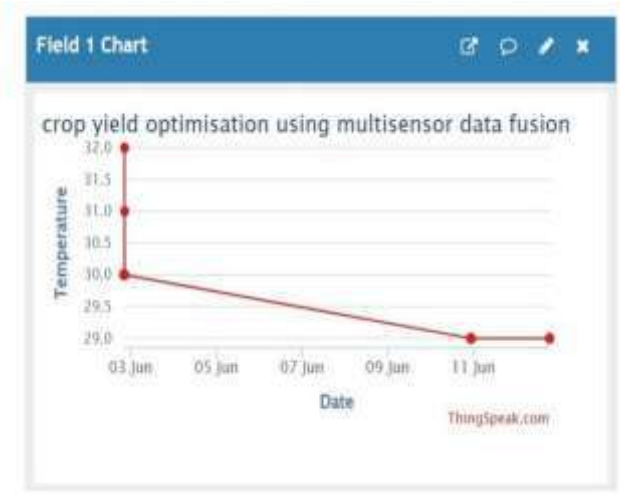


Figure 3: Graphical representation of temperature variation with date for crop yield optimization using multi sensor data fusion on the ThingSpeak platform.
Source: Authors, (2026).

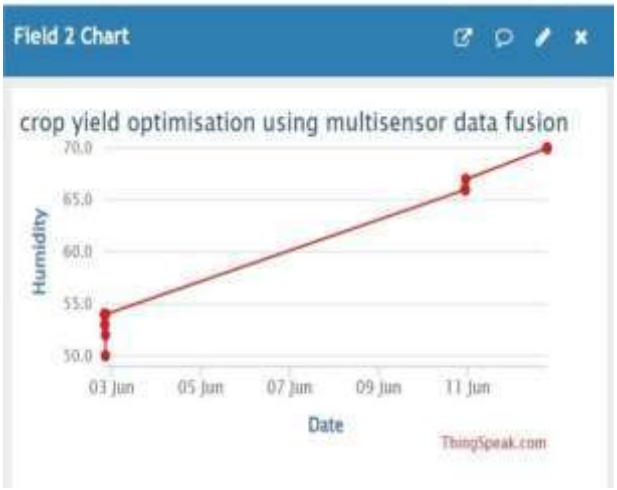


Figure 4: Graphical representation of Humidity and it's Variation of levels over time for crop yield optimization using multi sensor data fusion on the ThingSpeak platform.
Source: Authors, (2026).

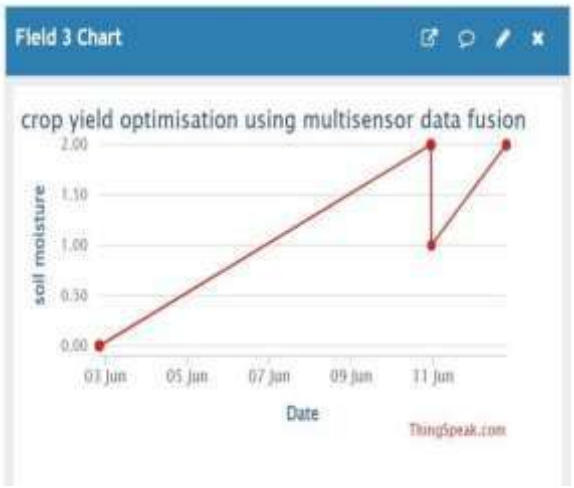


Figure 5: Graphical representation of soil moisture variation with date for crop yield optimization using multi sensor data fusion on the ThingSpeak platform.
Source: Authors, (2026).



Figure 6: Graphical representation of gas level variation with date for crop yield optimization using multi sensor data fusion on the ThingSpeak platform.
Source: Authors, (2026).

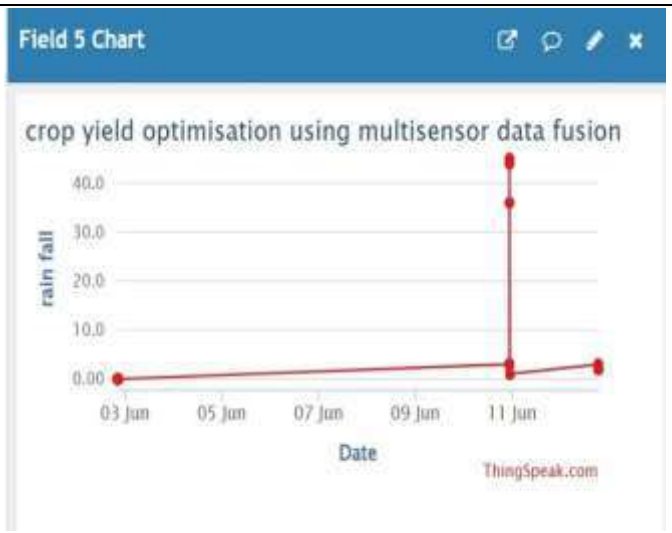


Figure 7: Graphical representation of rainfall variation with date for crop yield optimization using multi sensor data fusion on the ThingSpeak platform.
Source: Authors, (2026).

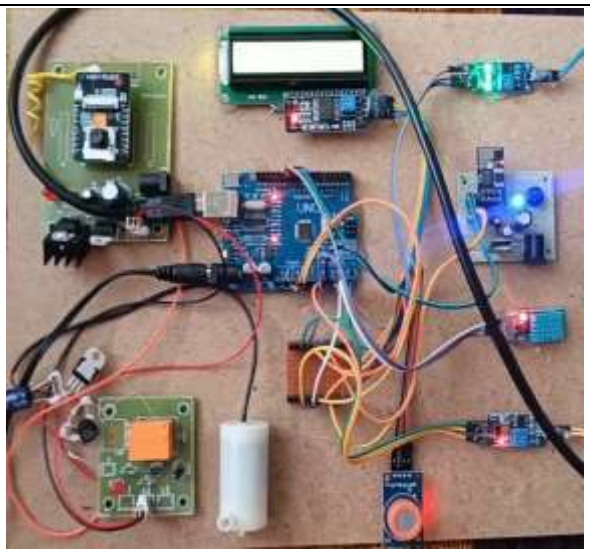


Figure 8: Prototype Outcome of IoT Based Multi sensor data fusion for Precision Crop yield Optimization.
Source: Authors, (2026).



Figure 9: Image analysis result showing leaf classification as good leaf using multi sensor data fusion with adjustable parameters such as resolution, brightness, and contrast.
Source: Authors, (2026).



Figure 10: Image analysis result showing leaf classification as bad leaf using multi sensor data fusion with adjustable parameters such as resolution.
Source: Authors, (2026).

V.CONCLUSION

This research effectively showcases the practicality and vast potential of an IoT-driven system designed for intelligent and holistic crop management. By seamlessly integrating real-time environmental monitoring with image-based crop health analysis, the system offers farmers a comprehensive, data-centric framework to enhance productivity and maintain healthy crops. The environmental monitoring unit, based on Arduino and ESP8266 modules linked with ThingSpeak, continuously collects environmental metrics such as temperature, humidity, rainfall, soil moisture, and CO_2 concentration—empowering farmers to make well-informed, data-backed decisions. Complementing this, the visual analysis unit, utilizing the ESP32-CAM, captures and processes leaf images using Embedded C algorithms to evaluate plant health.

A notable innovation in this model is its automation mechanism, where the system activates a submersible pump through a relay when unhealthy leaf conditions are detected, ensuring immediate corrective action. This integrated setup enhances crop protection, optimizes water and fertilizer usage, and significantly reduces manual labor, time, and costs. The edge-level image processing on the ESP32 ensures faster local responses, while ThingSpeak's cloud platform facilitates long-term data storage, visualization, and analysis. Together, these components form a scalable, affordable, and intelligent solution that empowers farmers to adopt precision agriculture techniques. Ultimately, this system contributes to improved crop vitality, higher yield efficiency, and sustainable farming practices through real-time data utilization and automation.

V.1 FUTURE SCOPE

The proposed IoT-based framework presents numerous opportunities for advancement.

Enhanced Disease/Pest Identification: Future developments can include advanced image recognition methods or AI-based models for specific disease or pest identification instead of simple “healthy” or “unhealthy” classification. Using extensive datasets and either cloud-based inference or optimized on-device AI models can enhance accuracy and detection capability.

Predictive Analytics and AI Integration: Incorporating predictive models that combine environmental and visual data could help forecast potential crop stress or diseases before they occur. For example, correlating low soil moisture with early wilting signs can enable timely interventions through AI-based analytics.

Multi-Actuator Control and Precision Application: The system can be scaled to control multiple actuators like automated nutrient delivery systems, misters, or robotic sprayers for precise pesticide application, promoting resource-efficient precision farming.
Mobile Application and User Interface Development: A user-friendly mobile app for Android and iOS platforms can be developed to provide real-time alerts, data visualization, remote control, and historical trend analysis, improving user interaction and decision-making.

Energy Optimization and Renewable Integration: To ensure continuous operation, especially in remote fields, power-efficient strategies such as solar-powered modules with energy storage can be introduced, minimizing energy dependency and enhancing sustainability.

Data Security and Privacy: Strengthening data security by using encrypted communication and secure login mechanisms will protect sensitive agricultural data on both ThingSpeak and local networks.
Integration with Agricultural Databases: Linking the system to agricultural knowledge bases can provide crop- specific insights, recommended environmental conditions, and best-practice treatments, enhancing the system’s intelligence and adaptability.

VI.AUTHOR’S CONTRIBUTION

Conceptualization: Chandraiahgari Dinesh Reddy and Danduprolu Kiran Kumar.
Methodology: Chandraiahgari Dinesh Reddy and Danduprolu Kiran Kumar.
Investigation: Chandraiahgari Dinesh Reddy and Danduprolu Kiran Kumar.
Discussion of results: Chandraiahgari Dinesh Reddy and Danduprolu Kiran Kumar.
Writing – Original Draft: Chandraiahgari Dinesh Reddy.
Writing – Review and Editing: Chandraiahgari Dinesh Reddy and Danduprolu Kiran Kumar.
Resources: Danduprolu Kiran Kumar.
Supervision: Chandraiahgari Dinesh Reddy and Danduprolu Kiran Kumar.
Approval of the final text: Chandraiahgari Dinesh Reddy and Danduprolu Kiran Kumar.

VII.ACKNOWLEDGMENT

The successful execution and completion of this project, "IoT-Based Multi-Sensor Data Fusion for Precision Crop Yield Optimization Using Arduino, ESP32, and Webcam Integration," was achieved through the unwavering support assistance of several contributors and organizations. We offer our profound appreciation to our academic institution and the faculty members who provided invaluable, continuous mentorship, technical direction, and crucial encouragement throughout every phase of the research and development process. Their expertise was instrumental in shaping the final methodology and outcomes of this work. Gratitude is also extended to the laboratory personnel and technical facilities which furnished the essential resources, operational equipment, and specialized workspaces required for the physical hardware implementation and rigorous data collection activities.

The authors further acknowledge the indispensable contribution of open-source tools and online platforms. Specifically, the functionalities provided by ThingSpeak and the Arduino Integrated Development Environment (IDE) were critical in enabling real-time data monitoring, cloud integration, and efficient analysis throughout this study. Finally, we recognize the foundational intellectual contributions made by the researchers and authors of the referenced studies. Their published works and insights provided the necessary conceptual framework and theoretical justification that supported the advancement of this current research.

VIII.REFERENCES

[1] A. Zakaria, A. Y. M. Shakaff, M. J. Masnan, F. S. A. Saad, A. H. Adom, M. N. Ahmad, M. N. Jaafar, A. H. Abdullah, and L. M. Kamarudin, "Improved maturity and ripeness classifications of Mangifera indica cv. Harumanis mangoes through sensor fusion of an electronic nose and acoustic sensor," Sensors, vol. 12, no. 5, pp. 6023–6048, May 2012. DOI: 10.3390/s120506023.

[2] H. Akcay, S. Kaya, E. Sertel, and U. Alganci, "Determination of olive trees with multi-sensor data fusion," in Proc. 8th Int. Conf. Agro-Geoinformatics (Agro-Geoinformatics), Jul. 2019, pp. 1–6. DOI: 10.1109/Agro-Geoinformatics.2019.8820464.

[3] J. Denize, L. Hubert-Moy, J. Betbeder, S. Corgne, J. Baudry, and E. Pottier, "Evaluation of using Sentinel-1 and-2 time-series to identify winter land use in agricultural landscapes," Remote Sens., vol. 11, no. 1, p. 37, Dec. 2018. DOI: 10.3390/rs11010037.

[4] J. Muangprathub et al., "IoT and agriculture data analysis for smart farm," Comput. Electron. Agricult., vol. 156, pp. 467-474, 2019. DOI: 10.1016/j.compag.2018.12.011.

[5] Y.-Y. Zheng, J.-L. Kong, X.-B. Jin, X.-Y. Wang, and M. Zuo, "CropDeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture," Sensors, vol. 19, no. 5, p. 1058, Mar. 2019. DOI: 10.3390/s19051058.

[6] B. Espejo-Garcia, J. Martinez-Guanter, M. Pérez-Ruiz, F. J. Lopez-Pellicer, and F. J. Zarazaga-Soria, "Machine learning for automatic rule classification of agricultural regulations: A case study in Spain," Comput. Electron. Agricult., vol. 150, pp. 343–352, Jul. 2018. DOI: 10.1016/j.compag.2018.05.023.

[7] S. R. Nandurkar, V. R. Thool, and R. C. Thool, "Design and development of precision agriculture system using wireless sensor network," in Proc. First Int. Conf. Automation, Control, Energy and Systems (ACES), 2014, pp. 1–6. DOI: 10.1109/ACES.2014.6808017.

[8] G. Sahitya, N. Balaji, and C. D. Naidu, "Wireless sensor network for smart agriculture," Int. J. Eng. Technol., vol. 8, no. 2, pp. 1109–1115, 2016. DOI: 10.21817/ijet/2016/v8i2/160802214.

[9] M. Ayaz, M. Ammad-Uddin, Z. Sharif, A. Mansour, and E. H. M. Aggoune, "Internet-of-Things (IoT)-based smart agriculture: Toward making the fields talk," IEEE Access, vol. 7, pp. 129551–129583, 2019. DOI: 10.1109/ACCESS.2019.2932609.

[10] R. S. Sinha, Y. Wei, and S. H. Hwang, "A survey on LPWAN technology of IoT for agriculture," J. ICT Standardization, vol. 5, no. 1, pp. 1–18, 2017. DOI: 10.13052/jicts2245-800X.511.

[11] D. Singh and J. S. Singh, "Soil moisture and temperature monitoring using IoT-based multi-sensor fusion," J. Agricult. Eng., vol. 57, no. 4, pp. 284–295, 2020. DOI: 10.37855/jae.2020.v57i04.001.

[12] N. R. Patel et al., "Data fusion of multiple sensors for crop yield optimization and soil health," Int. J. Agricult. Biol. Eng., vol. 11, no. 3, pp. 1–8, 2018. DOI: 10.25165/j.ijabe.20181103.3241.

[13] S. V. Zade et al., "Performance analysis of ESP32 and Arduino in real-time farming data acquisition," Int. J. Sci. Res. Sci. Technol., vol. 7, no. 3, pp. 220–225, 2020. DOI: 10.32628/IJSRST207361.

[14] A. Kumar and V. Kumar, "Design of IoT-based smart irrigation system using soil moisture sensor," in Proc. Int. Conf. I-SMAC (IoT in Social, Mobile, Analytics and Cloud), 2017, pp. 593–597. DOI: 10.1109/I-SMAC.2017.8058250.

[15] X. Wang, "Computer vision and webcam integration for real-time crop disease detection," Sensors, vol. 22, no. 4, p. 1530, 2022. DOI: 10.3390/s22041530.

[16] P. Devi and R. Sharma, "Multi-sensor data fusion for precision agriculture: A review," IEEE Sensors J., vol. 21, no. 11, pp. 1234–1245, 2021. DOI: 10.1109/JSEN.2021.3066123.

[17] M. S. Sirsat, E. Cernadas, M. Fernández-Delgado, and R. Khan, "Classification of agricultural soil parameters in India," Comput. Electron. Agricult., vol. 135, pp. 269–279, 2017. DOI: 10.1016/j.compag.2017.01.019.

[18] H. A. Orengo et al., "Automated detection of archaeological mounds using machine-learning classification of multisensor satellite data," Proc. Nat. Acad. Sci. USA, vol. 117, no. 31, pp. 18240–18250, 2020. DOI: 10.1073/pnas.2005583117.

[19] R. Katarya, A. Raturi, A. Mehndiratta, and A. Thapper, "Impact of machine learning techniques in precision agriculture," in Proc. 3rd Int. Conf. Emerg. Technol. Comput. Eng., 2020, pp. 1–6. DOI: 10.1109/ICETCE48199.2020.9091741.

[20] A. Khanna and S. Kaur, "Evolution of Internet of Things (IoT) and its significant impact in the field of Precision Agriculture," Comput. Electron. Agricult., vol. 157, pp. 218-231, 2019. DOI: 10.1016/j.compag.2018.12.039.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 612-619. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2903

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

COMPREHENSIVE ENERGY AUDIT AND EFFICIENCY ASSESSMENT OF A UNIVERSITY CAMPUS: A CASE STUDY AT THE UNIVERSITY OF EASTERN PHILIPPINES

Bless Ampuan*[1], Rem Laodeño [2] and Merewina Llanie Tapong[3]

[1, 2]Department of Mechanical Engineering, University of Eastern Philippines, 6400, Catarman, Philippines.
[3]Department of Agricultural and Biosystems Engineering, University of Eastern Philippines, 6400, Catarman, Philippines.

[1]http://orcid.org/0009-0009-2063-8566, [2]http://orcid.org/0009-0007-6952-2455, [3]http://orcid.org/0009-0006-2009-1431

Email: aelabuep@gmail.com, *bless.ampuan@uep.edu.ph, llanietapong5@uep.edu.ph

ARTICLE INFO

Article History
Received: November 7, 2025
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Energy audit;
Energy management;
Energy conservation measures (ECMs);
University energy consumption;
Sustainable campus operations

ABSTRACT

This study presents a comprehensive energy audit of the University of Eastern Philippines (UEP) aimed at assessing electricity consumption patterns, identifying inefficiencies, and proposing energy conservation strategies for institutional sustainability. The audit followed a three-phase framework—pre-audit planning, detailed audit and analysis, and post-audit evaluation—guided by the standards of the International Energy Agency (IEA) and ASHRAE. Data were collected from 25 major campus facilities using electricity billing records (2016–2018), field measurements, and stakeholder interviews. Results revealed a 51.4% increase in total electricity expenditure over three years, from ₱4.77 million to ₱7.22 million, driven by infrastructure expansion, increased equipment usage, and reliance on air-conditioning systems. The Cocofed Building, College of Business Administration (CBA) Building, and Bio-Physical Laboratory were identified as the highest energy-consuming structures, accounting for a substantial portion of total campus demand. Seasonal variations showed peak energy use from May to October, corresponding to high cooling loads. Based on audit findings, the implementation of targeted Energy Conservation Measures (ECMs)—including LED retrofitting, HVAC optimization, behavioral reforms, and potential solar photovoltaic integration—could achieve up to 15% reduction in energy use, equivalent to approximately ₱692,000 in annual savings. The study concludes that systematic energy auditing provides an essential framework for data-driven decision-making and long-term sustainability in higher education institutions, particularly in tropical and developing regions.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Electricity is a fundamental form of energy that underpins modern economic activities and daily life, enabling essential services such as lighting, heating, cooling, and the operation of appliances in residential, commercial, and institutional buildings. Its absence would impair the operation of systems and hinder even simple activities that depend on it [1]. Energy consumption in buildings represents one of the major contributors to global energy demand and greenhouse gas emissions. The building sector accounts for approximately 30% of the world's total final energy consumption and nearly 28% of energy-related carbon dioxide emissions, primarily due to heating, cooling, lighting, and appliance use [2]. In the Philippines, commercial and residential buildings collectively constitute a major source of energy use and related greenhouse gas emissions, with commercial buildings alone representing approximately 10% of total national energy consumption according to the DOE's Energy Efficiency and Conservation Action Plan. Buildings and structures contribute significantly to the country's emissions profile, and subnational inventories — such as those from Pasig and Parañaque — estimate that buildings can account for up to 37% of local greenhouse gas emissions, driven primarily by lighting, cooling, and appliance loads.

These findings align with global trends, where buildings represent roughly one-third of total final energy use and energy-related CO_2 emissions, reinforcing the importance of energy audit and efficiency measures in mitigating both energy consumption and carbon outputs [3], [4]. As developing economies expand, public institutions such as universities are experiencing a rapid increase in energy demand because of infrastructure growth, increased occupancy, and the integration of technology-intensive facilities [5], [6]. Consequently, energy audits have become a fundamental tool for identifying inefficiencies and implementing conservation measures in academic and institutional environments [7]. Studies show that energy consumption in university campuses is highly dependent on building type, occupancy schedule, and operational factors. A bibliometric analysis of energy uses in higher education institutions found that electricity — especially for cooling, lighting, and equipment — is the primary energy load in campus buildings, and that functional differences between academic, residential, and research facilities significantly influence consumption patterns [8]. Statistical modeling of building energy use further demonstrates that operational schedules and environmental conditions significantly influence consumption patterns in campus buildings [9]. Local Philippine studies confirm that building function affects load distribution, with air-conditioning systems dominating in office and library buildings and lighting and plug loads varying by facility type [10].

According to the International Energy Agency (IEA), global energy demand in the buildings sector — including educational facilities — continues to rise due to modernization of infrastructure, extended operational hours, and the growing integration of digital technologies, which increase electricity loads for computing, lighting, and environmental controls. Buildings now account for roughly one-third of final energy consumption worldwide, making energy efficiency improvements critical in educational institutions that operate complex facilities with varied end-uses [11]. Universities, as microcosms of urban systems, possess a diverse mix of energy loads — ranging from classrooms, dormitories, and administrative offices to laboratories, libraries, and research centers. Each facility typically exhibits distinct consumption patterns influenced by building function, occupancy schedules, equipment utilization, and climatic conditions, highlighting the need for campus-wide diagnostic assessment. International systematic reviews of higher education institutions confirm that building function and operational characteristics are key determinants of energy demand, with heating, ventilation and air conditioning (HVAC), lighting, and plug-load equipment driving energy use across mixed-use campus buildings [8]. Thus, institutional energy audits serve as an essential diagnostic process to systematically assess where, how, and why energy is consumed within a university campus, forming the basis for benchmarking, identifying priority efficiency measures, and informing decision-making toward sustainable campus energy management [12], [13].

At the University of Eastern Philippines (UEP), rising energy costs have become a major operational concern. Annual electricity expenditures increased from PHP 3.2 million in 2012 to over PHP 7.2 million in 2018, marking a 51.4% escalation over six years. This trend underscores the urgent need for strategic energy management to improve operational efficiency and reduce financial strain. By implementing a systematic energy audit and management framework, the institution can potentially achieve a 15% reduction in energy use, equivalent to approximately PHP 692,000 in savings—funds that could be reallocated to academic and infrastructure development. Recent research reinforces the importance of university-based energy audits in improving energy performance. A study [14] demonstrated that implementing solar-assisted retrofits and control measures in university residential buildings reduced electricity use by 21%. Similarly, [15] performed a comprehensive audit on Moroccan university facilities, identifying HVAC inefficiencies as the largest contributors to wastage. In Malaysia, [16] emphasized that energy efficiency management frameworks are critical to ensuring sustainable campus operations, while [10] developed a procedural model for Philippine academic buildings, highlighting the significance of systematic energy monitoring.

Laboratories and specialized research centers are among the most energy-intensive campus facilities, consuming five to ten times more energy per square meter than standard offices due to ventilation, lighting, and specialized equipment requirements [9]. Therefore, targeted auditing approaches that address equipment scheduling, process optimization, and environmental control systems are crucial for such facilities. As the study [13] and [17] argue, integrating audit data with management systems and policy reforms is essential to establishing long-term sustainability across higher education institutions. This study aims to conduct a comprehensive energy audit at the University of Eastern Philippines (UEP) to evaluate existing energy utilization patterns, identify high-consumption areas, and propose actionable conservation measures. Specifically, the research (1) inventories energy-consuming facilities, (2) assesses consumption trends across multiple years, and (3) recommends efficiency interventions supported by cost–benefit analysis. The overarching goal is to establish a data-driven foundation for an institutional energy management framework that can serve as a model for other public universities in the Philippines and comparable tropical regions. Ultimately, the findings contribute to the global discourse on energy-efficient educational campuses by demonstrating how systematic auditing can foster operational savings, sustainability, and environmental stewardship in developing contexts.

II. MATERIALS AND METHODS

II.1 RESEARCH DESIGN AND FRAMEWORK

This research adopted a mixed-method design integrating both quantitative and qualitative approaches to systematically evaluate energy consumption patterns at the University of Eastern Philippines (UEP). The methodological framework was anchored on energy auditing guidelines and practices adopted in the Philippines, drawing primarily from the Department of Energy (DOE–Philippines) Energy Efficiency and Conservation (EEC) guidelines, and was supplemented by internationally recognized standards such as those of the International Energy Agency (IEA) and the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). The selection of this approach was informed by previous Philippine-based studies on energy audits in higher education institutions, which demonstrated that a structured audit process combining baseline data collection, field measurements, and consumption analysis is effective in identifying major energy loads and efficiency opportunities in university buildings [10]. The study was implemented in three phases: (1) pre-audit planning and baseline data collection, (2) detailed energy audit and consumption analysis, and (3) post-audit evaluation and formulation of energy efficiency recommendations.

This phased approach is consistent with methodologies applied in local university energy audit studies, which emphasized the importance of systematic assessment and post-audit evaluation to ensure practical and policy-aligned energy conservation measures [18]. [19], [20]. Each phase followed a systematic and sequential process to ensure comprehensive assessment of the university's facilities and an accurate representation of its energy consumption behavior within the local institutional and regulatory context.

II.2 STUDY SITE AND SCOPE

The study was conducted at the University of Eastern Philippines (UEP), a state university located in Catarman, Northern Samar, Philippines. The university comprises several academic, administrative, residential, and auxiliary buildings, each with distinct energy consumption characteristics. A total of twenty-five major facilities were included in the energy audit, encompassing both academic and non-academic structures such as the Cocofed Building, College of Business Administration (CBA) Building, Bio-Physical Laboratory, UEP-RCCC & Criminology Building, Diagnostic Center, and Hostel. These buildings were selected based on their relative energy intensity and their substantial contribution to the total electricity expenditure of the university between 2016 and 2018. The scope of the study focused primarily on electricity consumption, as it represented the dominant component of institutional energy use.

II.3 PHASE I: PRE-AUDIT PLANNING AND DATA COLLECTION

The pre-audit phase established the foundation for the detailed audit through data gathering and coordination with relevant university units. Initial consultations were conducted with facility managers, deans, administrative officers, and maintenance staff to obtain operational insights regarding building functions, occupancy schedules, and typical energy usage patterns [7]. A series of walkthrough inspections were then performed across all selected buildings to observe lighting systems, air-conditioning units, laboratory and computing equipment, and other major electrical installations. During this phase, historical electricity billing records from 2016 to 2018 were obtained from the Accounting and Facilities Management Offices to establish the university's baseline energy consumption and cost profile. In addition to data collection, process flow and energy utility diagrams were developed to map the flow of electricity within individual buildings and across the campus. This initial phase provided critical information on the distribution of energy loads, enabling the research team to identify the most energy-intensive areas for further investigation [15]. The combination of documentary data, physical observation, and stakeholder interviews provided a holistic understanding of how energy was consumed and managed within the university.

II.4 PHASE II: DETAILED ENERGY AUDIT AND ANALYSIS

The second phase involved conducting a detailed energy audit following ASHRAE Level II standards, which emphasize the integration of utility data analysis, field measurements, and performance evaluation [16]. Real-time consumption data were gathered from sub-meters and smart monitoring devices installed in high-consumption buildings such as the Bio-Physical Laboratory and Cocofed Building. For structures without sub-metering, consumption estimates were derived using connected load analysis, considering equipment wattage, operational hours, and occupancy patterns [6]. The audit team used portable digital analyzers to measure electrical parameters, including voltage, current, and power factor, while lighting levels were measured using a lux meter to assess efficiency and illumination adequacy. HVAC performance was evaluated based on system runtime, thermostat settings, and coefficient of performance (COP). These measurements were complemented by direct observations to identify operational inefficiencies such as equipment left running during idle periods or inadequate temperature control. Data analysis focused on determining consumption trends, load distributions, and seasonal variations in energy use.

Monthly and annual consumption were computed and compared across buildings to identify "energy guzzlers" or facilities with disproportionate consumption relative to their size or function. The study also employed normalized performance indicators such as energy use intensity (kWh/m²/year) to facilitate comparison among different types of buildings. Statistical analysis was performed using descriptive and trend analysis techniques, consistent with methodologies applied in similar university audits [14], [15]. The results provided insights into the operational dynamics driving the university's increasing energy expenditures. Energy Conservation Measures (ECMs) were identified based on the audit findings, focusing on both technical and behavioral opportunities for improvement. The potential measures included upgrading lighting systems to energy-efficient LEDs, optimizing HVAC temperature settings, improving equipment maintenance, introducing occupancy sensors, and exploring renewable energy options such as solar photovoltaic installations [5], [10], [14]. Each proposed intervention underwent cost-benefit evaluation, estimating potential savings, implementation costs, and payback periods to determine their feasibility and prioritization [9].

II.5 PHASE III: POST-EVALUATION AND RECOMMENDATIONS

The final phase of the study involved consolidating the audit results into a comprehensive Energy Management Plan (EMP) for the university. This plan detailed the recommended interventions, ranked according to their cost-effectiveness, energy-saving potential, and implementation feasibility. Follow-up consultations were held with the university administration to validate the findings and ensure alignment with institutional priorities and budgetary capacities. The EMP also proposed the establishment of an Energy Management Committee responsible for monitoring the implementation of energy-saving measures, verifying their performance, and institutionalizing continuous improvement practices in line with sustainable campus operations.

II.6 DATA ANALYSIS TOOLS AND EVALUATION

Quantitative data analysis was conducted using Microsoft Excel and SPSS software to process numerical data, generate statistical summaries, and produce graphical representations of energy consumption trends. Correlation analyses were employed to identify relationships between energy use and factors such as building function, occupancy rate, and seasonal variation.

Qualitative data gathered from interviews and observations were coded thematically to identify recurring issues related to operational practices and equipment management. To ensure methodological rigor, triangulation was applied by cross-verifying information from billing data, field measurements, and personnel interviews, thereby enhancing the reliability and validity of the results [7], [13], [16].

II.7 ETHICAL AND ADMINISTRATIVE CONSIDERATIONS

Prior to data collection, formal approval was secured from the University Research Office and the Office of the Vice President for Research, Development, and Extension. All building data and energy records were treated as confidential and used solely for research purposes. The study also adhered to the ethical standards of institutional research, ensuring that participation from staff and personnel during interviews was voluntary and conducted with informed consent.

III. RESULTS AND DISCUSSION

III.1 OVERVIEW OF ENERGY CONSUMPTION TRENDS

As shown in Fig. 1, the total energy expenditure of the university increased significantly from 2016 to 2018, reflecting the growth in infrastructure and equipment usage. The analysis of the University of Eastern Philippines' (UEP) electricity records from 2016 to 2018 revealed a continuous rise in total energy consumption across the main campus. Annual energy expenditures increased from ₱4.77 million in 2016 to ₱7.22 million in 2018, representing a 51.4% growth over three years. This escalation mirrors global patterns observed in institutional and educational facilities, where energy use continues to rise due to campus expansion, higher equipment density, and intensified reliance on information technology [2], [6], [16].

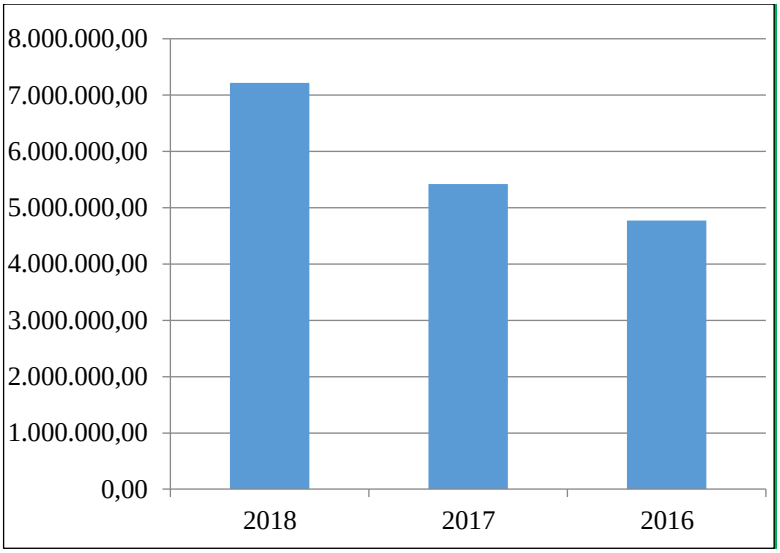


Figure 1: Annual Eelctricity Expenditure of the University of Eastern Philippines, 2016-2018.
Source: Authors, (2026).

Comparable trends were reported by [5] in Oman, who observed a 35% increase in energy demand in higher education institutions due to infrastructure expansion and increased cooling loads. Similarly, [15] found that Moroccan universities experienced significant increases in electricity consumption attributed to inefficient lighting and HVAC systems. These studies collectively affirm that academic institutions, particularly in tropical and developing regions, are facing substantial challenges in balancing modernization with energy efficiency. The growth of UEP's consumption can thus be attributed to the proliferation of electronic equipment, extended operational hours, and the installation of air-conditioning systems across academic and administrative spaces. A study [16] and [21] emphasized that these factors—combined with low awareness of energy conservation—are major contributors to energy inefficiency in Southeast Asian campuses.

III.2 BUILDING LEVEL ENERGY DISTRIBUTION

A building-by-building analysis revealed in Figure 2 that a few key facilities were responsible for a disproportionately large share of total electricity consumption. The Cocofed Building recorded the highest energy use, followed by the College of Business Administration (CBA) Building, Bio-Physical Laboratory, and UEP-RCCC & Criminology Building. Between 2016 and 2018, the Cocofed Building's annual cost increased from ₱1.099 million to ₱1.235 million, while the CBA Building's rose from ₱643,000 to ₱822,000.

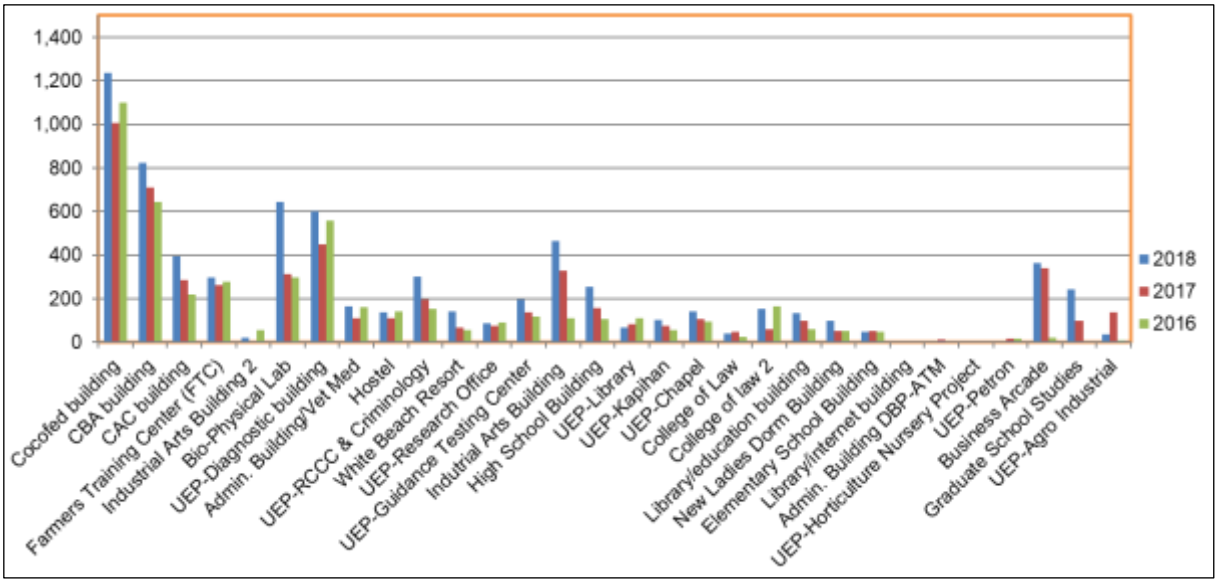


Figure 2: Comparative Annual Electricity Consumption by Building, 2016-2018.
Source: Authors, (2026).

This pattern is consistent with findings by [14], who identified that laboratory and administrative facilities exhibit the highest energy intensity due to equipment load, continuous operation, and greater cooling demands. Similarly, a recent benchmarking study conducted by [22] in Indonesia confirmed that laboratory buildings exhibited the highest Energy Consumption Intensity (IKE), exceeding 300 kWh/m² per year—significantly higher than classroom buildings. Furthermore, the unequal distribution of energy use across buildings aligns with the observations of [15], who found that just five university buildings in Morocco accounted for over 60% of total campus electricity use. This suggests that UEP's consumption profile—where a few buildings dominate total demand—is typical for higher education institutions, reinforcing the need for targeted, facility-specific energy interventions.

III.3 MONTHLY CONSUMPTION PATTERNS AND SEASONAL VARIATIONS

Monthly data analysis revealed cyclical consumption behavior influenced by both climatic conditions and the university's academic calendar. Electricity use peaked between May and October, coinciding with elevated outdoor temperatures and intensive academic activities. The Cocofed Building, for instance, recorded its highest monthly consumption in October 2018 (₱126,587.39), roughly double its January level. This trend highlights the strong relationship between tropical climatic conditions and increased HVAC load intensity. Studies by [14] and [23] report that cooling energy can account for up to 45% of total building energy consumption in humid regions. Similarly, research conducted at Nanjing Audit University in China [17] demonstrated that temperature-sensitive operations significantly contribute to monthly energy-use variability, particularly in facilities where ventilation and air-conditioning systems lack automated control mechanisms. In line with these findings, a pilot study [24] at the University of Eastern Philippines demonstrated the economic benefits of implementing automated electric fans, resulting in improved energy efficiency and reduced operational costs. Seasonal variations were also influenced by the academic schedule. Energy consumption typically declined during semester breaks, while examination and enrollment periods exhibited spikes in load due to extended operational hours. Such temporal fluctuations underscore the importance of load management strategies, including demand-side scheduling and programmable thermostats, as recommended by [16] and [10].

III.4 IDENTIFICATION OF HIGH-CONSUMPTION ZONES

The detailed audit revealed notable energy anomalies across several UEP buildings. The Bio-Physical Laboratory displayed monthly fluctuations between ₱26,762.05 in January and ₱76,211.21 in October, suggesting inconsistent operational management of laboratory equipment. Likewise, the Diagnostic Center showed a steady upward trend, doubling from ₱31,140 in January to ₱68,693 in December. Such variability is typical of research and medical facilities, where high plug loads associated with laboratory and medical equipment and stringent ventilation requirements contribute to substantial energy demand and operational losses. Laboratory fume hoods and high-level biosafety labs — common in research campuses — are known to be among the most energy-intensive building systems, often requiring constant air changes to maintain safety, which in turn increases HVAC energy use relative to conventional spaces. Similarly, medical facilities consume significant energy due to non-stop operation of equipment and environmental control systems, with refrigeration and specialized devices accounting for large shares of electricity use and continuous standby power draw.

Studies have shown that medical plug loads can constitute a significant portion of total electrical consumption in hospital laboratories, and that such equipment often exhibits high base-load and standby energy use even when not actively in operation, underscoring the importance of understanding load diversity in institutional energy audits [25], [26]. A 2022 study at Universiti Teknologi MARA (Malaysia) [27] reported that uncontrolled use of laboratory equipment and insufficient HVAC zoning led to 25% excess energy use annually. Similarly, [23] observed that without behavioral interventions, staff and student practices can negate the benefits of technical retrofits. The Industrial Arts Building exhibited a mid-year surge, likely linked to machinery-intensive workshops during training periods. These findings emphasize that operational scheduling, preventive maintenance, and behavioral modification are equally critical to energy efficiency as technological upgrades, a conclusion also reached by [7] and [22].

III.5 ENERGY CONSERVATION MEASURES (ECMs) AND EXPECTED IMPACT

Based on the audit results, several Energy Conservation Measures (ECMs) were proposed for implementation. These include replacing fluorescent and incandescent lamps with LED lighting, optimizing HVAC temperature settings, instituting preventive maintenance programs, installing occupancy sensors, and exploring on-site solar photovoltaic (PV) generation. LED retrofitting alone could yield 30–50% reductions in lighting energy use, while HVAC optimization could achieve an additional 15–20% reduction, consistent with results from previous university audits [5], [14], [15]. The integration of PV systems, particularly for high-load facilities like the Cocofed and CBA buildings, could further offset 10–15% of grid-supplied electricity, as demonstrated by [5] and [17]. A study [15] documented a 33% reduction in overall energy consumption following a comprehensive university retrofit in Morocco, while [17] achieved a 40% reduction through building-envelope improvements and smart control systems at Nanjing Audit University. These benchmarks suggest that UEP's targeted 15% reduction is realistic and achievable under existing operational and financial constraints. The successful implementation of ECMs would translate to annual savings of approximately ₱692,000, which could be reallocated to support campus development and research. Moreover, the institutionalization of an Energy Management Committee would facilitate long-term monitoring, verification, and sustainability, as recommended by [13].

III.6 COMPARATIVE ANALYSIS WITH OTHER INSTITUTIONS

When compared with other international energy audit studies, UEP's consumption profile and potential savings fall within a common range for tropical universities. [22] reported that Indonesian universities achieved 18–22% energy reductions after adopting targeted ECMs. In Malaysia, [16] and [21] documented up to 20% savings following the implementation of behavioral awareness programs and smart control technologies. Similarly, [17] in China achieved significant efficiency gains through green envelope retrofits and integrated energy management systems. These findings collectively indicate that universities operating in tropical and humid climates—like those in Southeast Asia—share common challenges related to air-conditioning dependence, inefficient equipment, and behavioral patterns. The consistency of UEP's results with these studies confirms that systematic energy audits are effective diagnostic tools for identifying both technical inefficiencies and policy gaps, ultimately leading to sustainable institutional energy governance.

III.7 SUMMARY OF FINDINGS

The comprehensive audit established that UEP's rising electricity consumption is driven by a combination of structural, operational, and behavioral factors. Buildings housing laboratories, administrative offices, and specialized facilities exhibited the highest energy intensities. Seasonal and academic-cycle variations further exacerbated load demands, particularly during the dry and hot months. By implementing cost-effective ECMs and adopting behavioral and policy-level strategies, the university could feasibly achieve a 15% reduction in energy use and align with international best practices in campus energy management. The study reinforces prior evidence that data-driven energy audits serve as essential foundations for developing institutional energy policies, improving operational efficiency, and advancing sustainability goals in public universities [6], [7], [15], [17], [22].

IV. CONCLUSIONS

This study conducted a comprehensive energy audit of the University of Eastern Philippines (UEP) to evaluate its electricity consumption patterns, identify inefficiencies, and propose data-driven conservation strategies. The analysis of three years of electricity records (2016–2018) revealed a 51.4% increase in energy expenditure, rising from ₱4.77 million to ₱7.22 million. This significant escalation highlights the growing energy demand associated with the expansion of academic infrastructure, increased equipment usage, and the intensified adoption of air-conditioning systems across campus. Building-level analysis showed that a small number of facilities—the Cocofed Building, College of Business Administration (CBA) Building, Bio-Physical Laboratory, and RCCC & Criminology Building—accounted for the majority of total energy consumption. These buildings exhibited high load densities due to continuous operations, specialized equipment, and extensive cooling requirements.

Monthly data indicated strong seasonal variation, with peak consumption occurring between May and October, corresponding to warmer weather conditions and heightened academic activity. The audit also uncovered several operational inefficiencies, including the use of outdated lighting systems, lack of zoning and control in HVAC units, and absence of automated energy management systems. By implementing targeted Energy Conservation Measures (ECMs) such as lighting retrofits, HVAC optimization, and behavioral reforms, UEP could feasibly reduce its electricity consumption by at least 15%, equivalent to an annual savings of approximately ₱692,000. Overall, the study demonstrates that systematic energy audits are essential tools for identifying both technical and behavioral inefficiencies in university environments. The findings provide a solid foundation for establishing a campus-wide energy management framework that promotes sustainability, reduces operational costs, and enhances institutional resilience against rising energy prices.

V. RECOMMENDATIONS

1. *Implement High-Impact Energy Conservation Measures (ECMs).* Immediate actions should prioritize cost-effective measures such as LED lighting retrofits, HVAC system maintenance, thermostat optimization, and installation of occupancy sensors. These interventions can yield substantial energy savings with minimal capital investment.
2. *Establish an Energy Management Committee (EMC).* A university-level EMC should be formed to coordinate and oversee all energy-related initiatives. This committee should develop policies, monitor energy performance, and ensure that efficiency measures are implemented and sustained over time.
3. *Adopt Smart Metering and Real-Time Monitoring Systems.* Installing sub-meters and digital energy monitoring tools will allow precise tracking of electricity usage per building, enabling early detection of anomalies, accurate benchmarking, and data-driven decision-making.

4. *Integrate Renewable Energy Technologies.* The university should explore the installation of solar photovoltaic (PV) systems on suitable rooftops, particularly high-load facilities such as the Cocofed and CBA buildings. Even partial adoption could offset a portion of grid-supplied electricity and demonstrate the university's commitment to sustainability.
5. *Strengthen Institutional Policies and Awareness Programs.* Energy efficiency should be embedded into university policies, procurement guidelines, and maintenance practices. Conducting awareness campaigns, student-led initiatives, and staff training can encourage responsible energy behavior and ensure long-term cultural change.
6. *Conduct Regular Energy Audits and Benchmarking.* The energy audit should be repeated every three to five years to evaluate progress and identify new opportunities for improvement. Establishing performance indicators such as energy consumption intensity (kWh/m²/year) will allow comparison with national and international benchmarks.
7. *Explore Long-Term Structural and Thermal Improvements.* Future infrastructure upgrades should consider energy-efficient building designs, thermal insulation, reflective roofing, and natural ventilation strategies to reduce dependency on mechanical cooling systems.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Rem Laodeño.
Methodology: Bless Ampuan*, Rem Laodeño and Merewina Llanie Tapong.
Investigation: Bless Ampuan*, Rem Laodeño and Merewina Llanie Tapong.
Discussion of results: Bless Ampuan.
Writing – Original Draft: Bless Ampuan.
Writing – Review and Editing: Bless Ampuan*, Rem Laodeño and Merewina Llanie Tapong.
Resources: Rem Laodeño.
Supervision: Bless Ampuan*, Rem Laodeño and Merewina Llanie Tapong.
Approval of the final text: Bless Ampuan*, Rem Laodeño and Merewina Llanie Tapong.

VII. ACKNOWLEDGMENTS

The researchers sincerely thank the University of Eastern Philippines (UEP) led by Dr. Cherry I. Ultra, for the financial support that made this study possible, and the University Administration, particularly the Office of the President and the Office of the Vice President for Research, Development and Extension, headed by Dr. Karina Cui-Lim, for their approval and support. Appreciation is also extended to the College of Engineering, the Facilities Management and Campus Development Office, the Accounting Office, the Electrical Services Division, and other university personnel for their cooperation and provision of essential data. The researchers likewise acknowledge the assistance of the Energy Audit Task Force, academic representatives, and administrative staff. Finally, heartfelt thanks are extended to the researchers' families, friends, and colleagues for their continued encouragement and support.

VIII. REFERENCES

[1] Bless Ampuan, Mechille Fabiala, J. L. Castillo, Jobert Jarogon, Y. B. Pantua, and R. A. Saluib, "ELECTRICAL SYSTEM ASSESSMENT OF BOARDING HOUSES IN UNIVERSITY OF EASTERN PHILIPPINES-MAIN CAMPUS: COMPLIANCE TO PHILIPPINE ELECTRICAL CODE," SSRN Electronic Journal, Jan. 2025, doi: https://doi.org/10.2139/ssrn.5286224.

[2] International Energy Agency, "Buildings – Energy System," IEA, 2023. [Online]. Available: https://www.iea.org/energy-system/buildings

[3] "Philippines Energy Efficiency and Conservation Action Plan 2016-2020 | ESCAP Policy Documents Management," policy.asiapacificenergy.org. https://policy.asiapacificenergy.org/node/2975

[4] "From emissions to efficiency: Embrace green building practices in the Philippines - Philippine Information Agency," Philippine Information Agency, Sep. 07, 2024. https://pia.gov.ph/press-release/from-emissions-to-efficiency-embrace-green-building-practices-in-the-philippines (accessed Jan. 16, 2026).

[5] S. A. Al Rashdi et al., "A case study on the electrical energy auditing and saving techniques in an educational institution (IMCO, Sohar, Oman)," Case Studies in Thermal Engineering, vol. 31, p. 101820, Mar. 2022, doi: https://doi.org/10.1016/j.csite.2022.101820.

[6] S. Johnsson, M. Andrei, and M. Johansson, "Harmonizing energy audit reporting: Addressing data loss and policy challenges in the EU member states," Energy, vol. 319, p. 135040, Feb. 2025, doi: https://doi.org/10.1016/j.energy.2025.135040.

[7] V. Liberova et al., "Unleashing Energy Potential: Insights of Energy Audit Practices," Energies, vol. 18, no. 3, pp. 522–522, Jan. 2025, doi: https://doi.org/10.3390/en18030522.

[8] J. P. Laporte and J. M. Cansino, "Energy Consumption in Higher Education Institutions: A Bibliometric Analysis Focused on Scientific Trends," Buildings, vol. 14, no. 2, p. 323, Feb. 2024, doi: https://doi.org/10.3390/buildings14020323.

[9] I. B. Benitez, K. I. Repedro, and T. Aung, "Forecasting Building Energy Consumption Using Statistical Models Incorporating Operational and Environmental Factors," 2024 19th International Conference on Emerging Technologies (ICET), pp. 1–6, Nov. 2024, doi: https://doi.org/10.1109/icet63392.2024.10935175.

[10] M. Chan, E. Arriola, R. Cuevas III, and J. G. Reyes, "DEVELOPMENT OF AN ENERGY AUDIT WORKING PROCEDURE FOR AN ACADEMIC UNIVERSITY OFFICE BUILDING IN THE PHILIPPINES," ASEAN Engineering Journal, vol. 11, no. 1, pp. 88–95, Mar. 2021, doi: https://doi.org/10.11113/aej.v11.16669.

[11] Energy Efficiency 2025 – Analysis - IEA, "Energy Efficiency 2025 – Analysis - IEA," IEA, Nov. 20, 2025. https://www.iea.org/reports/energy-efficiency-2025.

[12] Issa Chaer, Bertug Ozarisoy, Mubarak, S. Salari, and Ye Zhihui, "Energy efficiency in educational buildings: A systematic review of smart technology integration and occupant behaviour," Building and Environment, vol. 280, pp. 113132–113132, May 2025, doi: https://doi.org/10.1016/j.buildenv.2025.113132.

[13] J. J. John, A. P. Azodo, E. U. Bawa-Boyi, and F. C. Mezue, "Energy Auditing for University Energy Management: A Tool for Enhancing Sustainability," Advances in Science and Technology, vol. 160, pp. 227–244, 2025. doi: 10.4028/p-x2str1

[14] I. Bosu, H. Mahmoud, and H. Hassan, “Energy audit, techno-economic, and environmental assessment of integrating solar technologies for energy management in a university residential building: A case study,” Applied Energy, vol. 341, p. 121141, 2023. doi: 10.1016/j.apenergy.2023.121141

[15] Badr Ouhammou et al., “Energy saving potential diagnosis for Moroccan university campuses,” AIMS energy, vol. 11, no. 3, pp. 576–611, Jan. 2023, doi: https://doi.org/10.3934/energy.2023030.

[16] Shafini Mohd Shafie, A. Harits Nu'man, and N. N. A. N. Yusuf, “STRATEGY IN ENERGY EFFICIENCY MANAGEMENT: UNIVERSITY CAMPUS,” vol. 11, no. 5, pp. 310–313, Aug. 2021, doi: https://doi.org/10.32479/ijeep.11265.

[17] Yuan Li, Bijun Wang, Yue Fei, Green renovation of building envelopes in higher education institutions: study of energy-saving and carbon reduction at Nanjing audit university Mochou campus, International Journal of Low-Carbon Technologies, Volume 20, 2025, Pages 1057–1079, https://doi.org/10.1093/ijlct/ctaf058

[18] R. P. Peña, “An energy audit study of Cavite State University Indang, Cavite,” Animo Repository, 2022. https://animorepository.dlsu.edu.ph/etd_masteral/6074 (accessed Jan. 16, 2026).

[19] N. S. Lopez, J. Gonzaga, and L. A. G. Lim, “Energy audit and analysis of the electricity consumption of an educational building in the Philippines for smart consumption,” 2017IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), Dec. 2017, doi: https://doi.org/10.1109/hnicem.2017.8269505.

[20] V. Jay, V. Ylaya, and L. Malicay, “Assessment of Energy Savings Potentials at University in Lanao del Norte, Philippines,” International Research Journal of Advanced Engineering and Science, vol. 7, no. 2, pp. 22–28, 2022, Available: https://irjaes.com/wp-content/uploads/2022/03/IRJAES-V7N1P337Y22.pdf

[21] W. W. Choong, Y.F. Chong, L.S Ting (2012). Implementation of energy management key practices in Malaysian Universities.

[22] V. Faniama et al., “Energy Audit Based on Energy Consumption Intensity for Energy Conservation in University Buildings,” Jurnal Otomasi Kontrol dan Instrumentasi, vol. 16, no. 1, pp. 53–67, 2024, doi: https://doi.org/10.5614/joki.2024.16.1.6.

[23] J. Litardo, R. Hidalgo-Leon, and G. Soriano, “Energy Performance and Benchmarking for University Classrooms in Hot and Humid Climates,” Energies, vol. 14, no. 21, p. 7013, Oct. 2021, doi: https://doi.org/10.3390/en14217013.

[24] Bless Ampuan, “DEVELOPMENT OF AN AUTOMATIC ELECTRIC FAN,” SSRN Electronic Journal, Jan. 2025, doi: https://doi.org/10.2139/ssrn.5286246.

[25] N. M. A. Rahman, L. C. Haw, and A. Fazlizan, “A Literature Review of Naturally Ventilated Public Hospital Wards in Tropical Climate Countries for Thermal Comfort and Energy Saving Improvements,” Energies, vol. 14, no. 2, p. 435, Jan. 2021, doi: https://doi.org/10.3390/en14020435.

[26] N. Christiansen, M. Kaltschmitt, F. Dzukowski, and F. Isensee, “Electricity consumption of medical plug loads in hospital laboratories: Identification, evaluation, prediction and verification,” Energy and buildings, vol. 107, pp. 392–406, Nov. 2015, doi: https://doi.org/10.1016/j.enbuild.2015.08.022.

[27] N. A. Ahmad, M. H. Dzulkefli, and F. Abdul Rahim, “Assessing the Energy Performance of Public University in Malaysia by using Energy Conservation Measure (ECM): A Case Study of UiTM Tapah, Malaysia,” International Journal of Sustainable Construction Engineering and Technology, vol. 14, no. 3, Sep. 2023, doi: https://doi.org/10.30880/ijscet.2023.14.03.027.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 620-627. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2909

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

REVIEW OF EMERGING AI TECHNIQUES FOR GREEN HYDROGEN IN SUSTAINABLE ENERGY

Hiba Abdulkareem Saleh[1*], Rana Khalid Sabri[2], Omar Ibrahim Alsaif[3*]

[1]Department of Chemical and Petroleum Industries Techniques Engineering, Technical Engineering College, Northern Technical University, Mosul, Iraq.

[2]Department of Cybersecurity Engineering and Cloud Computing Technologies, Technical Engineering College for Computer and AI, Northern Technical University, Mosul, Iraq.

[3]Polytechnique College/Mosul, Northern Technical University, Mosul, Iraq.

[1]https://orcid.org/0009-0008-3533-527X, [2]https://orcid.org/0009-0009-1675-664X, [3]http://orcid.org/0000-0003-2832-7868

Email: *hiba.abdlkareem@ntu.edu.iq , rana.khalids@ntu.edu.iq, omar.alsaif@ntu.edu.iq

ARTICLE INFO

Article History
Received: November 8, 2025
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Green Hydrogen
Artificial Intelligence
Electrolyzer
Deep Learning
Machine Learning.

ABSTRACT

Green hydrogen is sustainable energy and clean source that has the potential to significantly alter important industries like manufacturing, the production of electricity, and transportation. Green hydrogen has the potential to ease the transition to low-carbon energy sources because of its leading role in global decarbonization initiatives. Increasing manufacturing productivity, ensuring the stability of storage, and coordinating usage strategies with commercial scalability remain significant challenges This review examines the application of artificial intelligence (AI) to these pressing issues throughout the green hydrogen value chain. In particular, it examines how optimization, deep learning, and machine learning (ML) can boost electrolyzer efficiency in fuel cell applications, improve energy conversion processes, forecast system performance, and optimize storage options. In addition, it is shown how AI-driven real-time monitoring systems and digital twins can be integrated into hydrogen infrastructures, showing how these tools make adaptive operational control and predictive maintenance easier. This article uses current technology advancements and a few selected case studies to critically examine how AI accelerates innovation in green hydrogen technologies in order to facilitate the development of affordable, scalable, and sustainable energy systems of the next generation.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The greenhouse gases, like methane, nitrous oxide and carbon dioxide, with the increase of emissions, have significantly exacerbated global warming. These gases cause irregularities in climate, threatening to raise sea levels, and subjecting the world to frequent natural disasters [1], [2]. Thus, transitioning to low-carbon, sustainable energy systems has become a strategic imperative in the world's search for immediate intervention to counter the challenging effects of climate change. Of these choices, green hydrogen is receiving attention as a promising vector of decarbonization for generating power, movement, and course industries [3]. Green hydrogen is the environmentally most sustainable type of hydrogen produced by electrolysis using renewable energy and emitting no greenhouse gases. Unlike grey, blue, or brown hydrogen produced using fossil fuels, they neither cause oxygen depletion nor increase water vapour. Yellow and white hydrogen share almost the same benefits, but are less scalable or occur naturally or in the rarest forms. Pink and turquoise hydrogen offer a moderate environmental impact, depending on whether one goes for the nuclear route or carbon-handling [4]. Figure. 1 display the production of green hydrogen types.

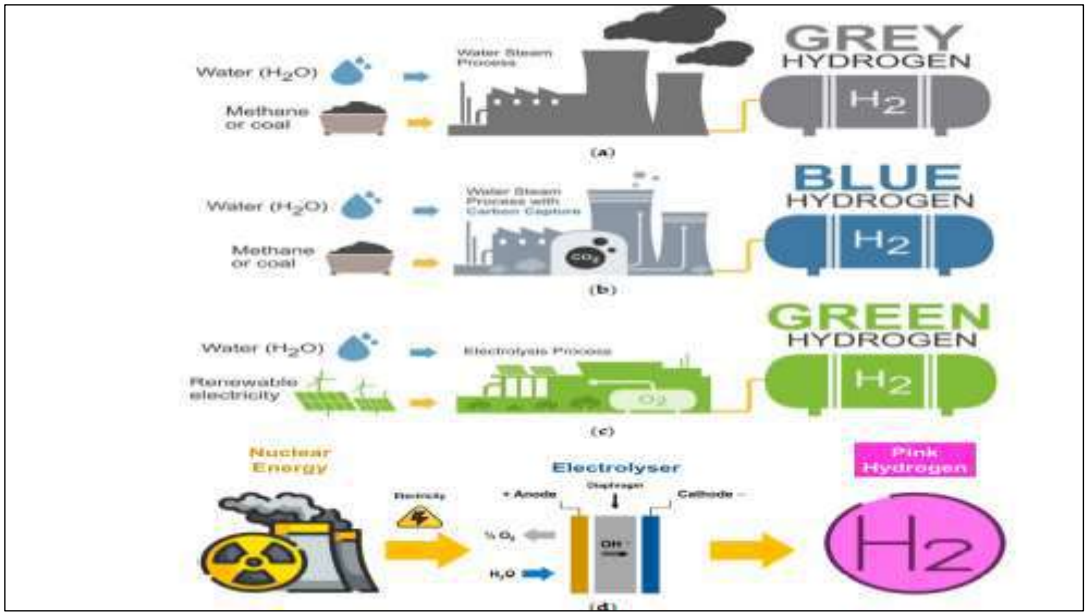


Figure 1: Production methods of hydrogen types.
Source: [5].

Green hydrogen is primarily an energy carrier that needs to remain clean and flexible, produced by splitting water molecules using electrolysis powered by wind, solar, or other renewable energy sources.at least slightly, or somewhat, zero emissions at the point of use [6], [7]. Electrolytic hydrogen production methods from renewable sources are compatible with modular and scalable infrastructures for energy generation, enabling constant flux to meet supply patterns and demand a capability that traditional fossil fuels cannot match [8]. While grey hydrogen uses steam methane reforming to produce (9–12) of CO_2 per ton of H_2, green hydrogen is carbon-free [9]. This flexibility is advantageous for ammonia production and other energy-intensive applications [10]. However, problems persist throughout the green hydrogen value chain despite these advantages. The two types (Metal hydrides and liquid organic hydrogen) carriers are storage technologies that are far behind conventional storage systems like underground caverns or compressed gas in terms of development, which results in inefficiencies and safety risks [11]. Moreover, the lack of an established distribution infrastructure and also technical challenges, such as hydrogen embrittlement and leakage caused by small molecular size, do not allow consideration of these highways to be integrated into existing natural gas grids [12].The term Artificial Intelligence (AI) is currently popular as a method to power solutions for these varying issues. Algorithms of optimization, machine learning, deep learning, and other methods of artificial intelligence can effectively model complex, nonlinear systems and draw valuable information from voluminous data [13].Figure. 2, illustrate how artificial intelligence and machine learning are infused into the hydrogen cycle, including production, storage, safety, and distribution, to produce a worthwhile energy system that is sustainable, safe, and efficient.

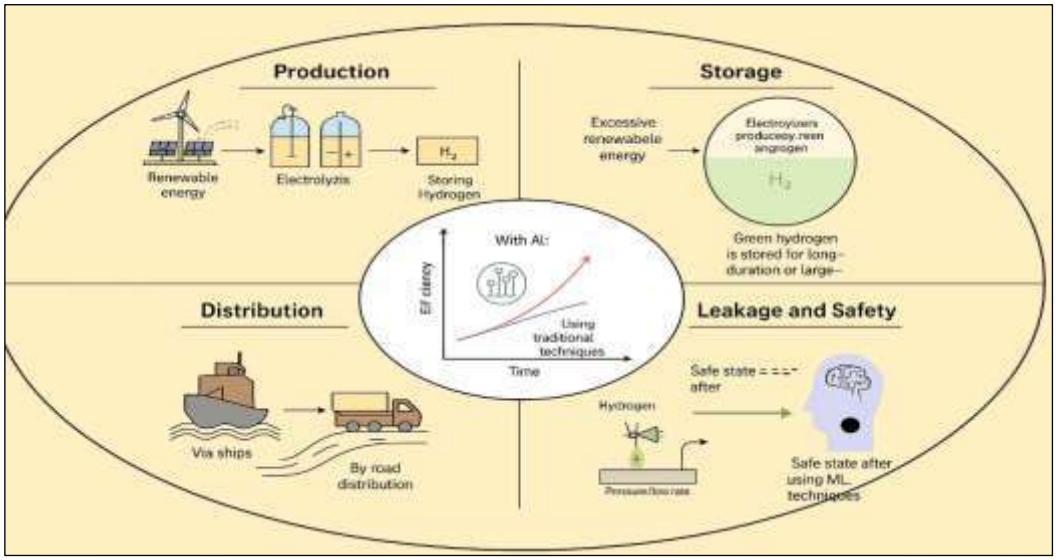


Figure 2: ProAI and ML integration across the hydrogen cycle from production to safety and distribution.
Source: [14].

Process control and reduce energy consumption in green hydrogen production, potentially increasing efficiency by up to 20%. Similarly, AI enables demand forecasting, predictive maintenance, and system-level integration with renewable energy grids for storage and utilization [15]. AI offers the disruptive path to accelerate the commercialization of green hydrogen technologies in response to the growing impetus for a global energy economy based on hydrogen.

II. GREEN HYDROGEN PRODUCTION USING RENEWABLE ENERGY SOURCES II.1 SUBTITLE

Green hydrogen production significantly reduces carbon emissions compared to conventional methods of hydrogen production by utilizing renewable energy from resources like solar ,wind power and other resources to drive water electrolysis. This sustainable strategy is necessary to advance clean energy and lessen the effects of greenhouse gases [16].Figure 3 demonstrates the integration of water electrolysis and energy storage with renewable energy sources like solar and wind energy in a grid-connected green hydrogen production facility.

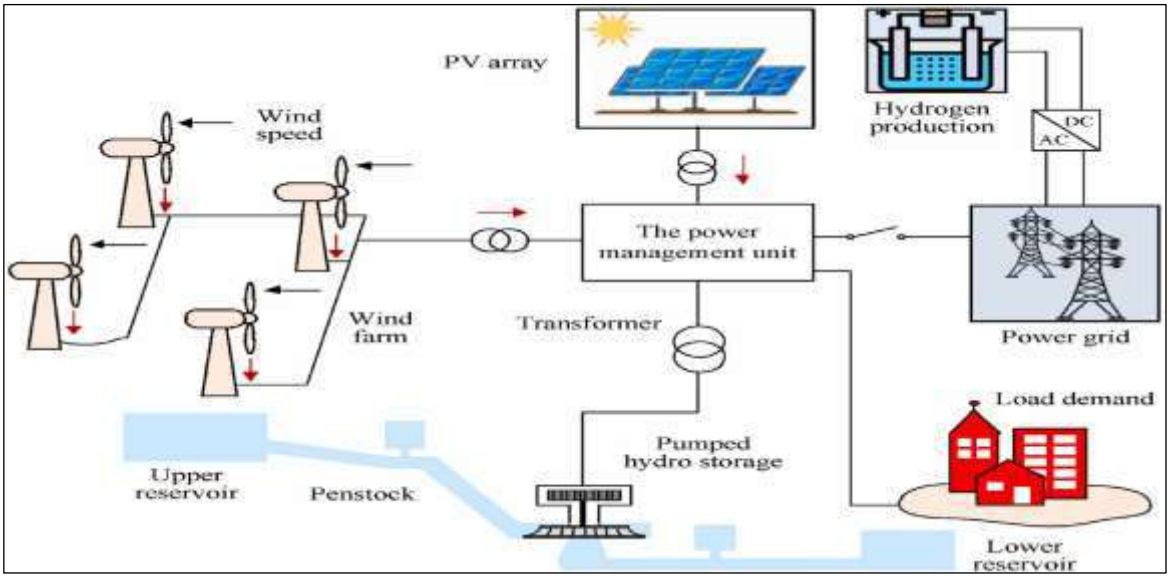


Figure 3: A diagram of a green hydrogen production system integrated into the grid using renewable resources, water electrolysis, and energy storage.
Source: [17].

The system can regulate fluctuations in the supply of renewable energy and maintain constant hydrogen production thanks to this integration. The testbed demonstrates, through its advanced power management and conversion technologies, how electrolysis and renewables can be balanced to create a more secure and environmentally friendly energy network. Examples of these developments include novel methods like the production of hydrogen through photo-electrochemistry and biological processes, as well as technological advancements in the production of green hydrogen such as enhanced electrolyzer performance thanks to cutting-edge catalysts and materials. These strategies also increase sustainability and efficiency by utilizing wind and solar power [18].In spite of the fact that technological advancements have proved the capability of photovoltaic (PV)-driven electrolysis systems to convert solar to hydrogen (STH) with high efficiency, there are still concerns regarding the stability of operations, safety, and storage under changing weather conditions [19]. By Combining different sources like(solar, wind, and geothermal energy) not only makes the system easier to scale up but also minimizes environmental impacts in a greater extent if backed by appropriate legal and economic frameworks [20].The national technoeconomic study has pointed that through using solar energy and various electrolysis technologies, Canada could theoretically produce more than 200 million tons of green hydrogen per year which would largely depend on regional optimization [21]. The 7 kW electrolysis system in Namibia has been successfully integrated with off-grid community wind energy and has shown that surplus wind energy can be utilized without affecting the system's load requirements to produce up to 235 kg of green hydrogen annually. This is a clean fuel substitute for the local needs that are not met like backup power and cooking [22]. In the same way, a regional study done in the Piura region estimated that yearly over 947,000 tons of green hydrogen produced through electrolysis which was majorly powered by solar PV (94.8%), with the rest wind and biomass. The study also put out that only 47.4% of hydrogen would be enough for the whole energy demand of the region and in this way there would be a significant reduction of greenhouse gas emissions [23].

III. GREEN HYDROGEN APPLICATIONS

Green hydrogen plays the role of hydrogen source in several applications such as clean fuel for transport, energy systems and industries that do not impact the environment. The most important areas in hydrogen utilization are mainly fuel cells, ammonia production, power-to-metal technology to produce aluminum and steel anodes from water electrolysis, and synthetic fuels. They all have a pivotal role in the development of global sustainability by consuming nearly half of the global production of hydrogen [24].Fuel cells are high-efficiency energy converters fueled by hydrogen, emitting no emissions. Adding membranes can produce additional pure water. Fuel cells make up 28% of final energy demand and are recommended by Peter Thiele of Siemens for managing electricity waste. Siemens fuel cells have an AC output of over 90%, making them an efficient option for powering various sectors globally [25].Energy systems, conversely, would stand to benefit greatly from green hydrogen, particularly when combined with intermittent renewables such as wind and solar. The storage of excess electricity as hydrogen and conversion back into electricity during times of demand dramatically improves grid stability and flexibility, and thereby serves, as a buffer against the variable nature of renewable generation [26] .The interest in green hydrogen for industrial applications that are hard to convert to renewable electricity is high. The amount of hydrogen needed is typically rather low, but is needed for specific purposes, including hydrogen as feedstock, and hydrogen as reducing agent for metal oxides, or synthesis of more complex molecules.

The high energy consumption, the specific application and the tradition of steam methane reforming lead to much of the industrial application being found at coal sites, steel locations or refineries, as well as in ammonia production for the fertilizer market [27] , [28].The iron and steel sector is the largest industrial consumer of hydrogen used in the reduction of iron ore pellets . In this process, hydrogen reacts with the oxidized concentrate and, in combination with the carbon oxide, generates a steel result. Still used in large quantities, carbon oxide is emitted to the atmosphere and will be limited to meet climate goals. One of the ways to decarbonise the steel process is to use green hydrogen generated with zero emissions [29],[30]. The production of chemicals, such as methanol or ammonia, is another hydrogen industrial application. In the ammonia production, steam methane reforming is the traditional process and the main world source, driven by its low capital cost, high efficiency and development in the first half of the 20th century. However, the use of steam methane reforming leads to a large hydrogen carbon footprint, in one of the sectors that most contributes to the global CO2 emissions [31]. Figure. 4 depicts current and future industrial applications of green hydrogen, emphasizing its importance in industries that include oil refining, ammonia, and methanol production, metalworking, and chemical processing.



Figure 4: Schematic diagram of a grid-connected green hydrogen production system consisting of renewable resources, water electrolysis, and energy storage.
Source: [32].

IV. ARTIFICIAL INTELLIGENCE ROLE IN GREEN HYDROGEN PRODUCTION TECHNOLOGIES

Artificial Intelligence plays a key role in driving green hydrogen production by making it more efficient and cost-effective. The use of machine learning on large datasets makes it possible to modify the operational settings of electrocatalysts making them perform better, produce more hydrogen, and consume less energy [33]. These artificial intelligence-based tools quickly match material compositions with desirable properties, powering the discovery of low-cost, high-performance electrode materials for water electrolysis. Load balancing, by storing the excess of possible solar and wind energy, is made easier through AI technologies because AI allows for dynamic adjustment of operations so that electrolyzers operate during peak generation only [34]. Refining demand-supply forecasting is a significant avenue of AI, which improves forecasts for renewable energy outturns and energy demand estimations, leading to better operation and innovation in green hydrogen systems [35]. Supported by data-driven screening and optimization, AI accelerates the design process of catalysts, which can be time-consuming and costly through traditional experiments, and enables pinpointing highly potent ones. With the help of a data-driven screening and optimization process, AI speeds up the design of catalysts, reduces experimental times and costs, and ensures the accurate discovery of high-performing materials [36]. Together, these advances highlight the importance of AI as a key enabler in the (sustainable, cheaper) development of green hydrogen technology .AI and automation are widely used in renewable energy, including solar and wind energy AI is the mainstay in the green hydrogen production process as it regulates the electrolyzer performance with the aid of cybernetic control[37].

The system has the ability to utilize weather forecasting to predict the renewable energy delivery and thus, to maintain the operation at the required level [38].AI has, in fact, become the key factor in the production of hydrogen and the co-generation of solar and wind power, both of which are working together to provide a more sustainable and energy-efficient system for the future. Energy was wisely handled and hence a significant reduction in carbon emissions was achieved which, in turn, is going to power a grid of the future that is going to be not only more reliable but also more sustainable [39]. Besides these, the AI application to predictive maintenance and the hydrogen production equipment life cycle extension was an advantage. AI will be deployed to analyze the data from sensors mounted on the trucks and simultaneously, an artificial intelligence system will be used to monitor, predict, and control the maintenance in a modern, efficient, and effective way. ITM Power has released a case study on its stage 2 and claimed an availability percentage that is up to 12% greater. This also worked out well for the equipment's lifespan as it increased by twice and operating costs got cut. These results indicate that AI can greatly improve hydrogen production industry wide in terms of both productivity and stability [40]. A new hybrid system for energy dangers and hydrogen generation has been named Zink-Zwischenschritt-Elektrolyseur (ZZE) in German. The model utilized several machine-learning models to predict the state of charge (SOC). The LSTM neural network achieved the lowest error in its forecasting (MAE = 3.21%). The precise SOC estimation is vital for the optimization of future ZZE management software.

The findings highlighted the support of data-driven AI methods in next-generation hydrogen storage technologies development [41]. Table 1 explain the summary of the key AI-centered works intended to enhance the range of green hydrogen production processes via electrolysis. It highlights that the new method can beat the traditional ones not only in terms of efficiency but also in accuracy of prediction and performance of systems in applications of Green Hydrogen Technologies.

Table 1: The AI-Driven Enhancement of Electrolysis-Based Green Hydrogen Production.

Study	Key focus	Model used	Key finding
[42]	AI in Electrolysis – Optimal Operational Prediction	Support Vector Machine (SVM), XGBoost, and Random Forest (RF) trained on electrolysis variables (electricity input, temperature, pressure, water quality). Dataset: experimental data from electrolysis processes	AI models predicted optimal electrolysis conditions, enhancing electrocatalyst performance and reducing energy loss. Outperformed manual tuning by ensuring consistent hydrogen output under varying inputs
[43]	Parametric Modeling of Alkaline Electrolyzers	MATLAB/Simulink parametric model using inputs like temperature, KOH concentration, and pressure. Not AI-based, serves as a traditional comparison baseline.	Higher temperatures and a 30 wt% KOH solution reduced cell voltage, thereby boosting system efficiency. Provides a baseline to compare improvements from AI-enhanced models
[44]	PEM(Proton Exchange Membrane) Electrolyzer Design Optimization via ML (Machine Learning)	Polynomial and Logistic Regression trained on 148 experimental samples to predict 11 PEM design parameters (e.g., electrode area, hydrogen production rate.	Achieved MAE of 0.615 and 83.6% classification accuracy. Enabled faster and more precise PEM cell design than conventional iterative approaches
[45]	Hybrid Modeling of Electrocatalysts	Adaptive Neuro-Fuzzy Inference System (ANFIS) + Pelican Optimization Algorithm (POA) using catalyst property datasets	A hybrid AI model improved hydrogen output by 1.7% vs traditional methods by optimizing catalyst behavior under electrolysis conditions.
[46]	DRL (Deep Reinforcement Learning) for Predictive Maintenance of Electrolysers	Deep Reinforcement Learning (DRL) with Long Short-Term Memory (LSTM) and feature selection. Dataset: sensor time-series data from electrolyser systems	DRL model identified sensor faults with 0.99 correlation and RMSE of 0.1351, enhancing maintenance efficiency and reducing downtime compared to manual diagnostics

Source: Authors, (2026).

V.ARTIFICIAL INTELLIGENCE TECHNIQUES FOR GREEN HYDROGEN STORAGE OPTIMIZATION

AI is a great help for green hydrogen storage. It helps to identify and control the production easily, monitor the status of storage, and predict the demand. It uses the same data analytics and optimization algorithms for efficiency and reliability thus solving the hydrogen storage and power supply cycle problem in a sustainable manner [47]. A smart system using neural networks can measure the non-linear dynamics of PEM (Proton Exchange Membrane) with a small average absolute prediction error of 0.05179. This makes the TITAN (Technology Intelligent Testbed for Advanced Networks) system more efficient and reliable for hydrogen generation and storage control compared to conventional methods of data collection. It is a revolution that not only accelerates the clean energy technology developments but also allows energy sustainability by supporting hydrogen energy developments through such innovations [48]. A MATLAB/Simulink model of the hydrogen subsystem was developed, employing a regression model to estimate hydrogen flow rates in response to input electrolyzer currents. A dynamic simulation of hydrogen storage is also developed, providing a reference for the high-pressure and compression pressures required and their effect on performance, as well as the evaluation of hydrogen storage, low energy density, and high pressure. This increases momentum for policy and regulatory support, as well as investment facilitation, for sustainable energy technologies. For this purpose, this work aims to integrate renewable-based energy storage solutions, enhance system reliability, and implement regulatory policies for sustainable energy solutions [49]. An AIoT-HES is developed based on a knowledge engineering model. With real-time analytics and predictive maintenance, hydrogen production, storage, and distribution costs are minimized while output is optimized. As a result, the AIoT ensures the operational security of the system by promptly identifying risks and determining intelligent storage conditions, which will be connected to other types of renewable energy. Maintenance or risk management of machine learning models can further enhance predictive capacity. The articles are also promoting sustainable energy alternatives, regulations, and a durable H2 infrastructure no information loss is incurred [50].

VI.APPLICATIONS OF AI IN GREEN HYDROGEN DEVELOPMENT ACROSS KEY COUNTRIES

Artificial Intelligence plays a crucial role in fronting the development of green hydrogen technologies, including production, storage, distribution, and utilization. Recent comprehensive reviews showing that AI techniques—such as Artificial Neural Networks (ANNs)—are largely employed in making electrolyzer technologies more efficient, forecasting inputs of renewable energy, optimizing storage procedures, and facilitating integration with a smart grid and GIS-based site selection [36].In India, a deep learning-based forecasting model for solar-powered green hydrogen production was introduced. The model works through four phases, including solar irradiance forecasting, electrolyzer scheduling, and cost-effective production of hydrogen through photovoltaic sources [51]. In Turkey, hybrid artificial intelligence modeling combining CEEMDAN and GRU neural networks for hourly hydrogen production prediction was developed for optimal scheduling and integration of hydrogen production within renewable energy mechanisms [52]. For both Saudi Arabia and Germany, AI modeling frameworks were developed for the estimation of green hydrogen generation by PtX systems. Those models consider data-driven forecasting and optimise the techno-economical feasibility of hydrogen generation from renewable sources. In Saudi Arabia, the high solar irradiance is used with AI to maximize electrolyzer performance and cost efficiency. In Germany, AI tools are used to manage intermittent wind and solar inputs while optimizing grid integration. The comparative analysis helps define strategies for the installation of green hydrogen infrastructure based on ambient and economic real-time parameters [53].

In the U.S., more than 500$ million in funding has been approved by the Department of Energy toward AI-assisted decarbonization and infrastructure development for hydrogen. According to the U.S. Energy Information Administration (EIA), AI applications optimizing electrolyzer performance and predictive maintenance would likely bring green hydrogen costs down by 20% or more in the next 10 years, further enhanced by heuristic algorithms optimizing operational parameters, yielding 15%-30% efficiency improvements in pilot projects by startups such as H2Pro. The mentioned optimization tools, besides various other AI applications such as predictive analytics, computer vision for leak detection, and AI-assisted planning, get federal tax credits and demonstration funds, which make hydrogen technologies near to commercialization [54].

VII. CONCLUSION

AI has an influential position to play in the full green hydrogen value chain transformation which means enhancing its efficiency, reliability, and scalability. All the way through production, storage, and use, AI is gradually taking over the solutions especially those which rely on machine learning, by modeling complex systems, optimizing performance, and interpreting high-dimensional data in a manner more superior than traditional physical modeling techniques. At the very start of the hydrogen production, the AI has shown its ability by synchronizing renewable electricity disruption with the hydrogen demand variations; thus the operation became more flexible and stable. The world's energy market is still moving to a point where hydrogen can be produced and used as a primary renewable source and, therefore, the traditional models will be unable to portray the complex and dynamic workflows of the new hydrogen economy. All these considered, the future of AI in this sector looks bright but at the same time, issues such as cybersecurity and system integrity are the major obstacles to widespread acceptance. In securing the digital infrastructure that is both safe and resilient, one should not forget that this has to happen in parallel with the technical advancements in the integration of AI in hydrogen systems. In spite of the extensive progress achieved in green hydrogen technologies, the world still has to deal with some economic and technical obstacles that are upholding the large-scale use of hydrogen all over the world. The main issues are the installation of pipelines and other necessary infrastructures that are very costly, the hard task of managing energy supply and demand because of the fluctuating nature of renewable sources, and the long-standing dependence on few and expensive catalyst materials. On top of that, a lack of technically skilled labor and the absence of detailed government laws and policies are making it even harder for the green hydrogen industry to expand gradually. Therefore, it can take to a long time for that to happen unless a well-coordinated approach is adopted consisting of combining technology development, policy intervention, and very long lasting investment plans to finally secure the transition to a low-carbon energy system that is equitable and mainly supported by green hydrogen.

VIII. AUTHOR'S CONTRIBUTION

Conceptualization: Hiba Abdulkareem Saleh , Rana Khalid Sabri , Omar Ibrahim Alsaif.
Methodology: Author One , Author Two and Author Three .
Investigation: Author One and Author Two.
Discussion of results: Author One, Author Two and Author Three.
Writing – Original Draft: Author One.
Writing – Review and Editing: Author One ,Author Two.
Resources: Author Two.
Supervision: Author Two and Author Three.
Approval of the final text: Author One, Author Two and Author Three.

IX. REFERENCES

[1] M. Filonchyk, M. P. Peterson, L. Zhang, V. Hurynovich, and Y. He, "Greenhouse gases emissions and global climate change: Examining the influence of CO2, CH4, and N2O," Science of The Total Environment, vol. 935, p. 173359, Jul. 2024, doi: 10.1016/j.scitotenv.2024.173359.

[2] J. Wang and W. Azam, "Natural resource scarcity, fossil fuel energy consumption, and total greenhouse gas emissions in top emitting countries," Geoscience Frontiers, vol. 15, no. 2, p. 101757, Mar. 2024, doi: 10.1016/j.gsf.2023.101757.

[3] M. J. B. Kabeyi and O. A. Olanrewaju, "Sustainable Energy Transition for Renewable and Low Carbon Grid E'lectricity Generation and Supply," Front. Energy Res., vol. 9, Mar. 2022, doi: 10.3389/fenrg.2021.743114

[4] A. Mohamed Elshafei and R. Mansour, "Green Hydrogen as a Potential Solution for Reducing Carbon Emissions: A Review," JENRR, vol. 13, no. 2, pp. 1–10, Feb. 2023, doi: 10.9734/jenrr/2023/v13i2257

[5] B. Łosiewicz, "Technology for Green Hydrogen Production: Desk Analysis," Energies, vol. 17, no. 17, p. 4514, Sep. 2024, doi: 10.3390/en17174514

[6] H. A. Saleh, O. I. Alsaif, and L. Y. Abdulkadir, "Adaptive Control of a DC Servo Motor Using Particle Swarm and Gray Wolf Optimization Algorithms," in AIP Conf. Proc., American Institute of Physics, 2025. doi: 10.1063/5.0260224.

[7] N. S. Hassan et al., "Recent review and evaluation of green hydrogen production via water electrolysis for a sustainable and clean energy society," International Journal of Hydrogen Energy, vol. 52, pp. 420–441, Jan. 2024, doi: 10.1016/j.ijhydene.2023.09.068.

[8] S. Ramakrishnan, M. Delpisheh, C. Convery, D. Niblett, M. Vinothkannan, and M. Mamlouk, "Offshore green hydrogen production from wind energy: Critical review and perspective," Renewable and Sustainable Energy Reviews, vol. 195, p. 114320, May 2024, doi: 10.1016/j.rser.2024.114320.

[9] R. Madurai Elavarasan, S. Afridhis, R. R. Vijayaraghavan, U. Subramaniam, and M. Nurunnabi, "SWOT analysis: A framework for comprehensive evaluation of drivers and barriers for renewable energy development in significant countries," Energy Reports, vol. 6, pp. 1838–1864, Nov. 2020, doi: 10.1016/j.egyr.2020.07.007

[10] M. Ouikhalfan, O. Lakbita, A. Delhali, A. H. Assen, and Y. Belmabkhout, "Toward Net-Zero Emission Fertilizers Industry: Greenhouse Gas Emission Analyses and Decarbonization Solutions," Energy Fuels, vol. 36, no. 8, pp. 4198–4223, Apr. 2022, doi: 10.1021/acs.energyfuels.2c00238

[11] D. Zhao et al., “A Green(er) World for A.I,” May 2022, pp. 742–750. doi: 10.1109/IPDPSW55747.2022.00126.

[12] I. A. Makaryan, I. V. Sedov, E. A. Salgansky, A. V. Arutyunov, and V. S. Arutyunov, “A Comprehensive Review on the Prospects of Using Hydrogen–Methane Blends: Challenges and Opportunities,” Energies, vol. 15, no. 6, p. 2265, Mar. 2022, doi: 10.3390/en15062265.

[13] H. Abdulkareem, O. I. Alsaif, L. Younis, and R. Khalid, “Performance Optimization of BLDC Motor Control Using Sand Cat Swarm Algorithm and Linear Quadratic Regulator,” J. Robot. Control., vol. 6, no. 1, pp. 306–314, 2025, doi: 10.18196/jrc.v6i1.24958.

[14] M. Motiramani et al., “AI-ML techniques for green hydrogen: A comprehensive review,” Next Energy, vol. 8, p. 100252, Jul. 2025, doi: 10.1016/j.nxener.2025.100252

[15] Doctoral Candidate, Department of Mechanical Engineering, Idaho State University, Idaho, USA et al., “AI-DRIVEN OPTIMIZATION IN RENEWABLE HYDROGEN PRODUCTION: A REVIEW,” ajis, vol. 6, no. 1, pp. 76–94, Apr. 2025, doi: 10.63125/06z40b13.

[16] S. G. Nnabuife, K. A. Quainoo, A. K. Hamzat, C. K. Darko, and C. K. Agyemang, “Innovative Strategies for Combining Solar and Wind Energy with Green Hydrogen Systems,” Applied Sciences, vol. 14, no. 21, Art. no. 21, Jan. 2024, doi: 10.3390/app14219771.

[17] S. G. Nnabuife, A. K. Hamzat, J. Whidborne, B. Kuang, and K. W. Jenkins, “Integration of renewable energy sources in tandem with electrolysis: A technology review for green hydrogen production,” International Journal of Hydrogen Energy, vol. 107, pp. 218–240, Mar. 2025, doi: 10.1016/j.ijhydene.2024.06.342.

[18] R. K. Sabri, L. Y. Abdulkadir, A. M. Khidhir, and H. A. Saleh, “Diagnosing Gingiva Disease Using Artificial Intelligence Techniques,” Diyala. J. Eng. Sci., vol. 18, no. 2, pp. 179–189, 2025, doi: 10.24237/djes.2025.18211.

[19]M. A. Abdelkareem et al., “Optimized solar photovoltaic-powered green hydrogen: Current status, recent advancements, and barriers,” Solar Energy, vol. 265, p. 112072, Nov. 2023, doi: 10.1016/j.solener.2023.112072.

[20] Q. Hassan, S. Algburi, A. Z. Sameen, H. M. Salman, and A. K. Al-Jiboory, “A review of green hydrogen production by renewable resources,” Energy Harvesting and Systems, vol. 11, no. 1, Jan. 2024, doi: 10.1515/ehs-2022-0127.

[21] G. K. Karayel and I. Dincer, “Green hydrogen production potential of Canada with solar energy,” Renewable Energy, vol. 221, p. 119766, Feb. 2024, doi: 10.1016/j.renene.2023.119766.

[22] S. Iileka, D. S. Likius, P. Tomas, D. K. Chembe, and H. MuAshekele, “Viability of Green Hydrogen Production at Benguela Wind Energy Community Project in Lüderitz, Namibia,” Journal of Sustainable Development of Energy, Water and Environment Systems, vol. 12, no. 3, Art. no. 1120498, Sep. 2024, doi: 10.13044/j.sdewes.d12.0498

[23] D. Caravantes et al., “Estimation of Green Hydrogen Production Potential in the Piura Region from Water Electrolysis,” in Proc. 2022 IEEE ICA-ACCA, Oct. 2022, pp. 1–6, doi: 10.1109/ICA-ACCA56767.2022.10006002.

[24]A. Yavari, C. J. Harrison, S. A. Gorji, and M. Shafiei, “Hydrogen 4.0: A Cyber–Physical System for Renewable Hydrogen Energy Plants,” Sensors, vol. 24, no. 10, p. 3239, Jan. 2024, doi: 10.3390/s24103239.

[25] E. Papagiannidis, I. M. Enholm, C. Dremel, P. Mikalef, and J. Krogstie, “Toward AI Governance: Identifying Best Practices and Potential Barriers and Outcomes,” Inf Syst Front, vol. 25, no. 1, pp. 123–141, Feb. 2023, doi: 10.1007/s10796-022-10251-y.

[26] U. Y. Qazi, “Future of Hydrogen as an Alternative Fuel for Next-Generation Industrial Applications; Challenges and Expected Opportunities,” Energies, vol. 15, no. 13, p. 4741, Jan. 2022, doi: 10.3390/en15134741.

[27] S. Damman, E. Sandberg, E. Rosenberg, P. Pisciella, and U. Johansen, Largescale hydrogen production in Norway - possible transition pathways towards 2050. SINTEF Digital, 2020. Accessed: Aug. 25, 2024. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2649737

[28] R. K. Sabri, L. Y. Abdulkadir, A. M. Khidhir, and H. A. Saleh, “Diagnosing Gingiva Disease Using Artificial Intelligence Techniques,” Diyala. J. Eng. Sci., vol. 18, no. 2, pp. 179–189, 2025, doi: 10.24237/djes.2025.18211.

[29] A. Bhaskar, M. Assadi, and H. Nikpey Somehsaraei, “Decarbonization of the Iron and Steel Industry with Direct Reduction of Iron Ore with Green Hydrogen,” Energies, vol. 13, no. 3, Art. no. 3, Jan. 2020, doi: 10.3390/en13030758.

[30] L. Ren, S. Zhou, and X. Ou, “The carbon reduction potential of hydrogen in the low carbon transition of the iron and steel industry: The case of China,” Renewable and Sustainable Energy Reviews, vol. 171, p. 113026, Jan. 2023, doi: 10.1016/j.rser.2022.113026.

[31] A. O. Oni, T. Giwa, C. Font-Palma, and D. A. Fadare, “Comparative techno-economic and life cycle greenhouse gas assessment of ammonia production from thermal decomposition of methane and steam methane reforming technologies,” International Journal of Greenhouse Gas Control, vol. 123, p. 103819, Feb. 2023, doi: 10.1016/j.ijggc.2022.103819.

[32] U. Y. Qazi, “Future of Hydrogen as an Alternative Fuel for Next-Generation Industrial Applications; Challenges and Expected Opportunities,” Energies, vol. 15, no. 13, p. 4741, Jun. 2022, doi: 10.3390/en15134741

[33] Z. Niu et al., “Generative Artificial Intelligence for Designing Multi-Scale Hydrogen Fuel Cell Catalyst Layer Nanostructures,” ACS Nano, vol. 18, no. 31, pp. 20504–20517, Aug. 2024, doi: 10.1021/acsnano.4c04943.

[34] M. Jamali, N. Hajialigol, and A. Fattahi, “An insight into the application and progress of artificial intelligence in the hydrogen production industry: A review,” Materials Today Sustainability, vol. 30, p. 101098, Jun. 2025, doi: 10.1016/j.mtsust.2025.101098.

[35] N. Zubairu, L. Al Jabri, and A. Rejeb, “A review of hydrogen energy in renewable energy supply chain finance,” Discov Sustain, vol. 6, no. 1, p. 186, Mar. 2025, doi: 10.1007/s43621-025-01007-0.

[36] R. Ding, J. Chen, Y. Chen, J. Liu, Y. Bando, and X. Wang, “Unlocking the potential: machine learning applications in electrocatalyst design for electrochemical hydrogen energy transformation,” Chemical Society Reviews, vol. 53, no. 23, pp. 11390–11461, 2024, doi: 10.1039/D4CS00844H.

[37] A. Abiola, F. S. Manzano, and J. M. Andújar, “A Novel Deep Reinforcement Learning (DRL) Algorithm to Apply Artificial Intelligence-Based Maintenance in Electrolysers,” Algorithms, vol. 16, no. 12, Art. no. 12, Dec. 2023, doi: 10.3390/a16120541

[38] K. Ukoba, K. O. Olatunji, E. Adeoye, T.-C. Jen, and D. M. Madyira, “Optimizing renewable energy systems through artificial intelligence: Review and future prospects,” Energy & Environment, vol. 35, no. 7, pp. 3833–3879, Nov. 2024, doi: 10.1177/0958305X241256293.

[39] B. B. Singla and N. S. Thakur, “Energy Transition through AI-Enabled Green Hydrogen Smart Grids,” vol. 7, no. 6, 2024

[40] A. K. Saxena, “AI-Driven Optimization for Green Hydrogen Production Efficiency,” 2024.

[41] D. Vila, E. Hornberger, and C. Toigo, “Machine learning based state-of-charge prediction of electrochemical green hydrogen production: Zink-Zwischenschritt-Elektrolyseur (ZZE),” Energy and AI, vol. 16, p. 100355, May 2024, doi: 10.1016/j.egyai.2024.100355.

[42] K. E. Bassey and C. Ibegbulam, “MACHINE LEARNING FOR GREEN HYDROGEN PRODUCTION,” Computer Science & IT Research Journal, vol. 4, no. 3, Art. no. 3, Dec. 2023, doi: 10.51594/csitrj.v4i3.1253.

[43] S. Niroula, C. Chaudhary, A. Subedi, and B. S. Thapa, “Parametric Modelling and Optimization of Alkaline Electrolyzer for the Production of Green Hydrogen,” IOP Conf. Ser.: Mater. Sci. Eng., vol. 1279, no. 1, p. 012005, Mar. 2023, doi: 10.1088/1757-899X/1279/1/012005.

[44] A. Mohamed, H. Ibrahem, R. Yang, and K. Kim, “Optimization of Proton Exchange Membrane Electrolyzer Cell Design Using Machine Learning,” Energies, vol. 15, no. 18, Art. no. 18, Jan. 2022, doi: 10.3390/en15186657.

[45] H. Rezk, A. G. Olabi, M. A. Abdelkareem, A. Alahmer, and E. T. Sayed, “Maximizing Green Hydrogen Production from Water Electrocatalysis: Modeling and Optimization,” Journal of Marine Science and Engineering, vol. 11, no. 3, Art. no. 3, Mar. 2023, doi: 10.3390/jmse11030617

[46] A. A. Alhussan et al., “Green hydrogen production ensemble forecasting based on hybrid dynamic optimization algorithm,” Front. Energy Res., vol. 11, p. 1221006, Jun. 2023, doi: 10.3389/fenrg.2023.1221006.

[47] L. Y. Abdulkadir, H. A. Saleh, O. I. Alsaif, and R. K. Sabri, “Evaluating Facial Emotional Proportion Based on Computer Vision Technique,” J. Imag. Gr., vol. 13, no. 5, pp. 469–475, 2025, doi: 10.18178/joig.13.5.469-475.

[48] S. A. Mohamed, O. I. Alsaif, and I. A. Saleh, “Intrusion Detection Network Attacks Based on Whale Optimization Algorithm,” Ing. Syst. Inf., vol. 27, no. 3, pp. 441–446, 2022, doi: 10.18280/isi.270310.

[49] Itani, Noureddine, and Nazih Moubayed. "Simulink Regression Modelling and Analysis for Production and Storage of Green Hydrogen." 2023 6th International Conference on Renewable Energy for Developing Countries (REDEC), pp. 7-12, IEEE, 2023,doi: 10.1109/REDEC58286.2023.10208170

[50] S. R. Joshua, S. Park, and K. Kwon, "Knowledge-Based Modeling Approach: A Schematic Design of Artificial Intelligence of Things (AIoT) for Hydrogen Energy System," in 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, Jan. 2024, pp. 235–241. doi: 10.1109/CCWC60891.2024.10427681.

[51] Bhuiyan, M. I. H., et al. (2024). Artificial Intelligence Applications for Hydrogen Energy Systems: A Comprehensive Review. Hydrogen, 5(6), 558–589. https://doi.org/10.3390/hydrogen5060032

[52] K. Sareen, B. K. Panigrahi, T. Shikhola, and R. Nagdeve, “Deep learning solar forecasting for green hydrogen production in India: A case study,” International Journal of Hydrogen Energy, vol. 50, pp. 334–351, Jan. 2024, doi: 10.1016/j.ijhydene.2023.08.323.

[53] K. Alhadhrami, A. Albalawi, S. Hasan, and A. M. Elshurafa, “Modeling green hydrogen production using power-to-x: Saudi and German contexts,” International Journal of Hydrogen Energy, vol. 64, pp. 1040–1051, Apr. 2024, doi: 10.1016/j.ijhydene.2024.03.161.

[54] K. Taghizad-Tavana, M. Ghanbari-Ghalehjoughi, A. Safari, M. T. Hagh, and A. E. Nezhad, “From green hydrogen production to artificial intelligence–driven energy management in hydrogen fuel cell electric vehicles: a comprehensive review of technologies, optimization techniques, international standards, and investment programs,” Applied Energy, vol. 399, p. 126534, Dec. 2025, doi: 10.1016/j.apenergy.2025.126534.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 628-636. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2920

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

ENHANCED CONTROL OF DFIG-BASED WIND TURBINE USING HYBRID FUZZY SUPER-TWISTING SLIDING MODE CONTROL FOR IMPROVED POWER QUALITY AND ROBUSTNESS

Anis Feddaoui[1], Lotfi Farah[2], Abdelouahab Benretem[3], Mohammed Abdeldjalil Djehaf[4], Youcef Djeriri[5]

[1, 2, 3] Genie Electromechanical Laboratory, Department of Electromechanical Badji Mokhtar University of Annaba, Algeria.
[4, 5] Faculty of Electrical Engineering, Djillali Liabes University Sidi Bel Abbes, Algeria.

[1]http://orcid.org/0009-0005-7679-6855, [2]http://orcid.org/0000-0002-5101-9698, [3]http://orcid.org/0009-0001-3155-7923
[4]http://orcid.org/0000-0002-4482-6918, [5]http://orcid.org/0000-0001-7590-8492

E-mail: anisfeddaoui.98@gmail.com, farahlot@yahoo.com, benretem_a@yahoo.fr, med_djehaf@yahoo.com, djeriri_youcef@yahoo.fr

ARTICLE INFO

Article History
Received: November 10, 2025
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
System wind turbine,
Doubly fed induction generator,
Fuzzy logic control,
Super-twisting sliding mode control,
Hybrid fuzzy super-twisting sliding mode control.

ABSTRACT

In the pursuit of advancing renewable energy technologies, optimizing wind turbine performance through sophisticated control systems is pivotal for addressing energy variability and enhancing system efficiency. This paper presents a novel hybrid fuzzy super-twisting sliding mode control (HFSTSMC) strategy for Doubly Fed Induction Generators (DFIGs) used in modern wind turbine configurations. DFIGs are favored for their adaptability and efficiency in managing variable wind speeds, making them integral to sustainable energy solutions. The proposed HFSTSMC integrates the robustness of super-twisting sliding mode control (STSMC) with the adaptive capabilities of fuzzy logic control (FLC), aiming to mitigate the chattering effect inherent in traditional sliding mode controls while maintaining high tracking accuracy and robustness. The hybrid controller utilizes fuzzy logic to dynamically adjust control gains based on system states, enhancing the dynamic performance and stability of DFIG-based wind turbines. Extensive modeling and simulation using MATLAB/SIMULINK validate the effectiveness of HFSTSMC in improving power quality, reducing mechanical stresses, and optimizing energy capture. The results demonstrate significant improvements in system response to rapid wind condition changes and overall operational efficiency under transient and uncertain conditions. This research underscores the substantial benefits of implementing HFSTSMC, highlighting its potential to revolutionize wind turbine control systems and contribute significantly to the global transition towards reliable and efficient renewable energy sources.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

As the global demand for renewable energy sources intensifies, wind energy has emerged as a vital component of sustainable energy portfolios [1]. The efficiency and reliability of wind turbines are paramount for harnessing this resource effectively. Doubly Fed Induction Generators (DFIGs) are widely employed in modern wind turbine systems due to their superior ability to handle variable wind speeds and contribute to the stability of electrical grids [2]. Traditional control strategies for DFIG-based wind turbines often face challenges such as chattering phenomena and sensitivity to system uncertainties, which can compromise performance and reliability. Sliding Mode Control (SMC) has been recognized for its robustness and effectiveness in dealing with non-linearities and uncertainties. However, conventional SMC techniques are prone to chattering, which can lead to excessive mechanical stress and energy losses [3]. [4]. To address these challenges, this paper proposes a novel Hybrid Fuzzy Super-Twisting Sliding Mode Control (HFSTSMC) approach.

This advanced control strategy combines the robustness of Super-Twisting Sliding Mode Control (STSMC) with the adaptability of Fuzzy Logic Control (FLC). The integration of fuzzy logic allows for dynamic adjustment of control parameters, thereby enhancing system performance and mitigating chattering effects[4-7]. The proposed HFSTSMC strategy aims to improve the dynamic performance, power quality, and operational stability of DFIG-based wind turbines. Extensive modeling and simulation studies are conducted using MATLAB/SIMULINK to evaluate the effectiveness of the proposed control approach. The results demonstrate significant improvements in system responsiveness to rapid changes in wind conditions, reduction in mechanical stresses, and overall optimization of energy capture. This research highlights the potential of HFSTSMC to revolutionize wind turbine control systems, offering a robust and efficient solution for maximizing wind energy utilization. The findings underscore the importance of advanced control strategies in the global transition towards renewable energy sources and the sustainable development of wind energy technology.

II. SYSTEM MODELING

II.1 AERODYNAMIC MODELING

The rotation of the blades due to wind speed generates mechanical power in the shaft. This mechanical power is then converted into electrical power by the Doubly Fed Induction Generator (DFIG)[8-10].

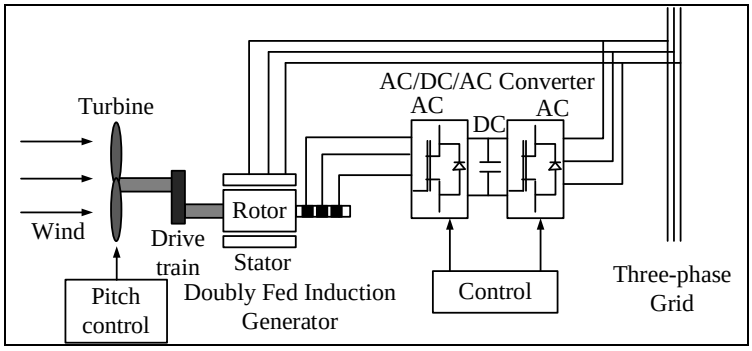


Figure 1: Wind power system.
Source: Authors, (2026).

The power output of the turbine can be described by Eq.(1):

$$P_T = \frac{1}{2} C_p(\lambda, \beta) \rho \pi R^2 {V_w}^3 \tag{1}$$

The power coefficient $C_p(\lambda, \beta)$ is determined using the equations provided in [1]:

$$C_p(\lambda, \beta) = \mathrm{C}_1 \left(\frac{C_2}{\lambda_i} - C_3 \beta - C_4 \right) e^{\frac{\mathrm{C}_5}{\lambda_\mathrm{i}}} + C_6 \lambda \tag{2}$$

With $\frac{1}{\lambda_\mathrm{i}} = \frac{1}{\lambda + 0.08\beta} - \frac{0.035}{\beta^3 + 1}$, $C_1 = 0.52$, $C_2 = 116$, $C_3 = 0.4$, $C_4 = 5$, $C_5 = 21$, $C_6 = 0.0068$ and $\lambda = \frac{R\omega_{mech}}{v_w}$

By maintaining the tip speed ratio λ at its optimal value (λ_{opt}) the coefficient is at its maximum then as shown in figure (2),the optimal power can calculate using the following equation :

$$P_s^* = \frac{1}{2} C_{pmax}(\lambda_{opt}) \rho \pi R^2 V_w^3 \tag{3}$$

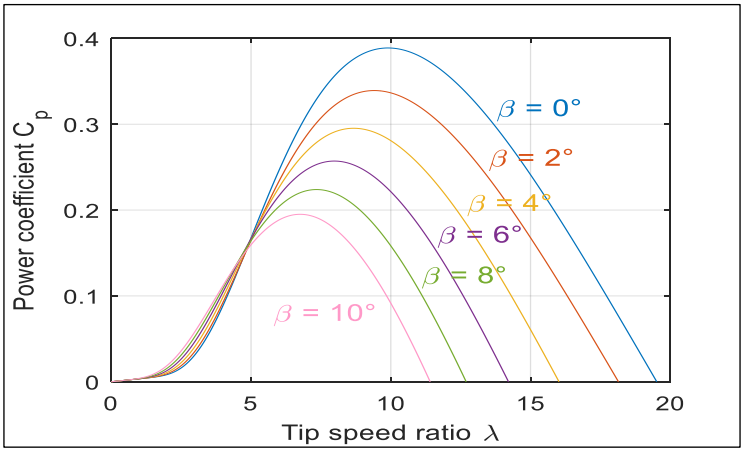


Figure 2: tip speed ratio (λ) vs. Power coefficient (Cp) curve.
Source: Authors, (2026).

II.2 DFIG SYSTEM MODELING

The Doubly Fed Induction Generator (DFIG) is a variable-speed wind turbine generator that efficiently converts wind energy into electrical power. DFIG employs both rotor and grid-side converters, allowing for better control of generator speed and enhanced power quality in the generated output [6-7].

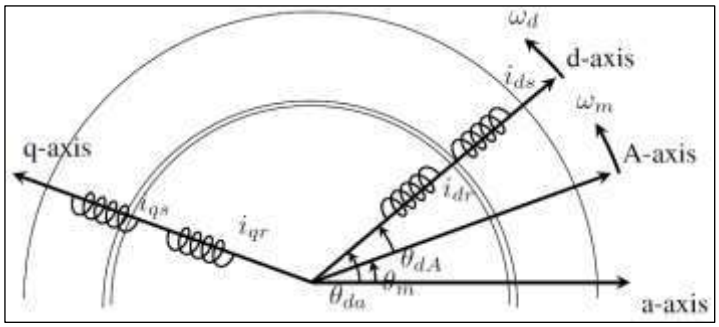


Figure 3: Illustration of stator and rotor with equivalent dq winding.
Source: [6].

After applying the Park transformation and subsequent mathematical derivation as referenced in [8], it was determined:

– The stator, rotor voltages

$$\begin{cases} v_{ds} = R_s i_{ds} + \frac{d}{dt}\phi_{ds} - \omega_d \phi_{qs} \\ v_{qs} = R_s i_{qs} + \frac{d}{dt}\phi_{qs} + \omega_d \phi_{ds} \\ v_{dr} = R_r i_{dr} + \frac{d}{dt}\phi_{dr} - \omega_{dA} \phi_{qr} \\ v_{qr} = R_r i_{qr} + \frac{d}{dt}\phi_{qr} + \omega_{dA} \phi_{dr} \end{cases} \quad (4)$$

– The fluxes can be found in Eq.(5)

$$\begin{cases} \phi_{ds} = L_s i_{ds} + L_m i_{dr} \\ \phi_{qs} = L_s i_{qs} + L_m i_{qr} \\ \phi_{dr} = L_r i_{dr} + L_m i_{ds} \\ \phi_{qr} = L_r i_{qr} + L_m i_{qs} \end{cases} \quad (5)$$

– The stator powers are presented in Eq.(6)

$$\begin{cases} P_s = v_{ds} i_{ds} + v_{qs} i_{qs} \\ Q_s = v_{qs} i_{ds} - v_{ds} i_{qs} \end{cases} \quad (6)$$

– The electromagnetic torque equation

$$T_{em} = \frac{P}{2} L_m \left(\phi_{qr} i_{dr} - \phi_{dr} i_{qr}\right) \quad (7)$$

From Eq.(4), by replacing the fluxes with rotor currents and after Laplace transformation, we obtain:

$$\begin{cases} i_{dr} = \frac{1}{R_r + S\sigma L_r} v'_{dr} \\ i_{qr} = \frac{1}{R_r + S\sigma L_r} v'_{qr} \end{cases} \quad (8)$$

Table 1: Parameters of wind turbine & DFIG.

Parameter	Definition	Value	Unit
J_{Turb}	Wind Turbine Moment of Inertia	2.4x10^6	Kg.m^2
ρ	Air Density	1.2	Kg.m^3
D	Rotor Diameter for the Wind Turbine (2R)	70.5	M
R_r	Resistance of the Rotor	1.5	$m\Omega$
R_s	Resistance of the Stator	2.0	$m\Omega$
L_r	Leakage of the Rotor	24	mH
L_s	Leakage of the Stator	24	mH
L_m	Mutual Leakage	23	mH
f_{syn}	Rated Frequency	60	Hz
P	Poles	6	.
J	DFIG Moment of inertia	75	Kg.m^2

Source: Authors, (2026).

III. CONTROLLER

III.1 SUPER TWISTING SLIDING MODE CONTROL

Super-Twisting Sliding Mode Control (STSMC) is a robust control technique that enhances the traditional SMC by reducing the chattering effect without compromising the system's robustness. The STSMC is a second-order sliding mode control method that does not require the derivative of the sliding variable, making it suitable for practical applications [9-13].

The STSMC law can be described by the following equations:

$$u = u_{eq} + u_n \tag{9}$$

Designing the sliding mode surface as [9] was the initial stage in creating the Super-Twisting Sliding Mode Controller (STSMC):

$$S = \left(\lambda + \frac{d}{dt}\right)^{r-1} e \tag{10}$$

The tracking error is represented by e, the system order is indicated by r, and the variable λ needs to meet the Hurwitz condition. According to Eq. (8), since r=1, our surface become:

$$S = e = I_{dqrDesired} - I_{dr} \tag{11}$$

The Lyapunov function is defined as:

$$V = \frac{1}{2}S^2 \tag{12}$$

The Lyapunov function derivative in Eq. (12) must meet $\dot{V} < 0$ as $S\dot{S} < 0$ in order to guarantee the stability criterion.
The sliding mode function's derivative is:

$$\dot{S} = \dot{e} = \dot{I}_{dqrDesired} - \dot{I_{dr}} \tag{13}$$

$$\dot{S} = \dot{e} = \dot{I}_{dqrDesired} + \frac{(R_r I_{dqr} - V_{dqr})}{\sigma L_r} \tag{14}$$

Therefore, the Lyapunov function may be derived as follows:

$$S\dot{S} = S\left(\dot{I}_{dqrDesired} + \frac{(R_r I_{dqr} - V_{dqr})}{\sigma L_r}\right) \tag{15}$$

To satisfy the condition $S\dot{S} < 0$, the sliding mode controller is designed as:

$$u_{dqr}^{eq} = \sigma L_r\left(\dot{I}_{dqrDesired} + R_r I_{dqr}\right) \tag{16}$$

Ultimately, a discontinuous control term has been introduced to determine the sliding mode control law, as illustrated in Eq.(17) :

$$u_n = -k_1\sqrt{|S|}sign(S) + \int(-k_2 sign(S))\, dt \tag{17}$$

Where k_1, k_2 are strictly positive gains.
Consequently, the resulting control signal is formulated as follow:

$$u = u_{eq} + u_n = \sigma L_r\left(\dot{I}_{dqrDesired} + R_r I_{dqr}\right) - k_1\sqrt{|S|}sign(S) + \int(-k_2 sign(S))\, dt \tag{18}$$

III.2 FUZZY LOGIC ENHANCEMENT

To enhance the performance of the STSMC, a fuzzy logic controller (FLC) can be incorporated to dynamically adjust the gains k_1 and k_2. The FLC uses system state information to modify these gains in real-time, improving the controller's adaptability and reducing residual chattering [14-16]. Input fuzzy membership functions, Our Fuzzy Logic Controller (FLC) uses input membership functions to establish a connection between the control decisions and the error signal. In Figure 4 [1], these membership functions are shown.

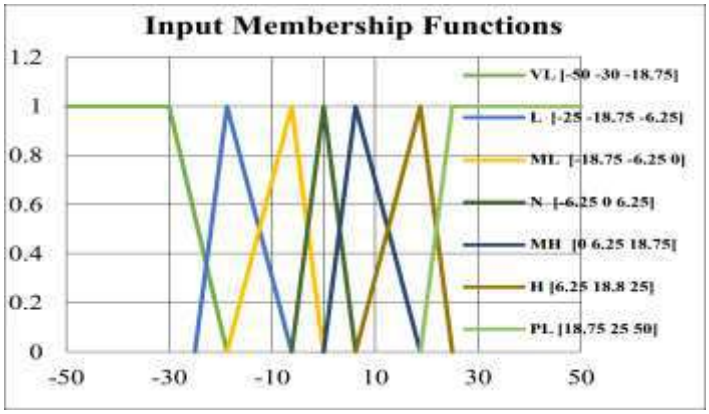


Figure 4: Input variables.
Source: Authors, (2026).

The linguistic variables are: "Very Low" ,"Low", "Moderately Low","Neutral","Moderately High","High" and "Very High" error magnitudes.

– Output fuzzy membership functions

The FLC can precisely modify the gains k_1 and k_2 of the Super-Twisting Sliding Mode Controller (STSMC) due to these output membership functions. This improves the overall performance and stability of the DFIG-based wind turbine by allowing it to function at its best in a variety of wind situations.

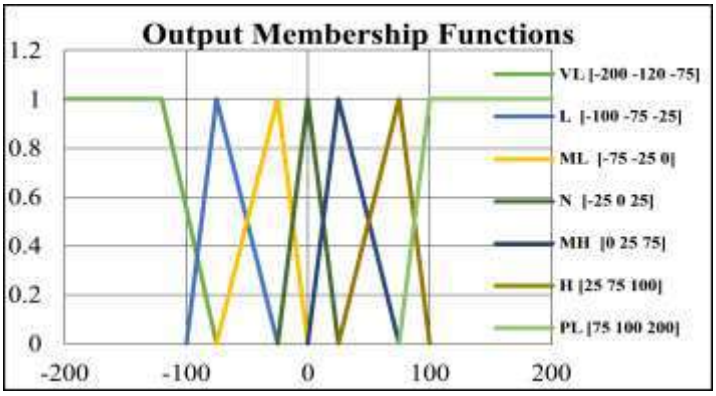


Figure 5: Output variables.
Source: Authors, (2026).

– Rules

These rules define how the FLC interprets input conditions and determines appropriate control actions, enabling it to mimic human-like reasoning and adapt to complex control scenarios.

Table 2: Fuzzy rules.

Output Input	VL	L	ML	N	MH	H	VH
VL	✔	.	.	.	.	.	.
L	.	✔	.	.	.	.	.
ML	.	.	✔	.	.	.	.
N	.	.	.	✔	.	.	.
MH	.	.	.	.	✔	.	.
H	.	.	.	.	.	✔	.
VH	.	.	.	.	.	.	✔

Source: Authors, (2026).

To create HFSTSMC, we replace the discontinuous sign function Sign(S) with a continuous fuzzy logic function Fuzzy(S). This modification helps to smooth out the control action and reduce chattering. The fuzzy logic controller dynamically adjusts the control signal based on the system's state.

The modified control law becomes:

$$u = \sigma L_r\left(\dot{I}_{dqrDesired} + R_r I_{dqr}\right) - k_1\sqrt{|S|}Fuzzy(S) + \int\left(-k_2 Fuzzy(S)\right)dt \quad (19)$$

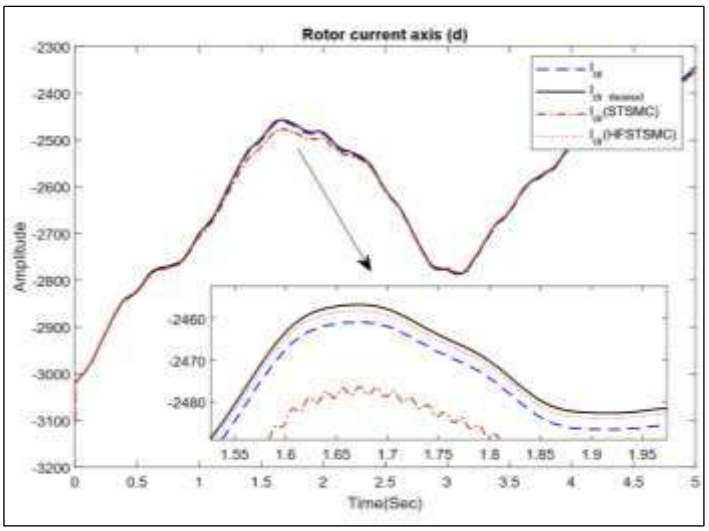
IV. RESULTS AND DISCUSSIONS

A Matlab simulation environment was used to assess the efficiency of the examined control techniques [17-19]. The main topics of the simulation were parameter fluctuation and reference tracking. The research made use of a Doubly Fed Induction Generator (DFIG) running at a nominal system voltage of Vs = 398/690V, with parameters shown in Table 1.

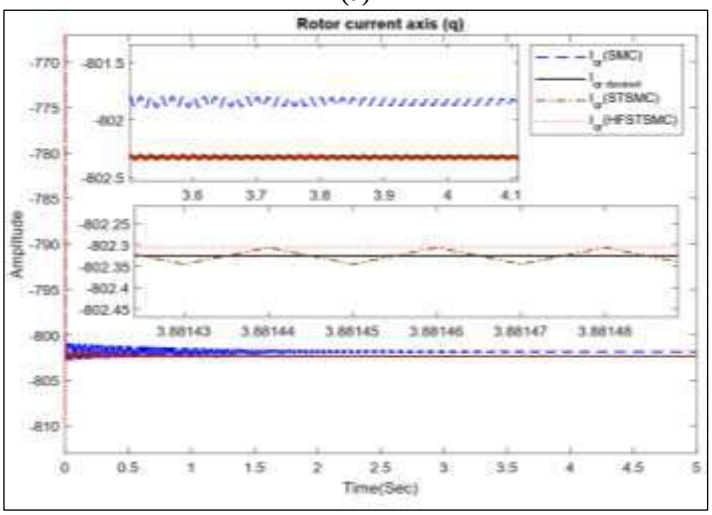
IV.1 TRACKING THE REFERENCE SIGNAL

The reaction of the Hybrid Fuzzy Super Twisting Sliding Mode Controller (HFSTSMC) to changing conditions in the environment was tested using a realistic wind speed profile. In order to maintain a unit power factor, the reference for reactive power (Qsref = 0) was set to zero and the reference for active power (Psref) was added using the MPPT technique. The performance of the Hybrid Fuzzy Super-Twisting Sliding Mode Control (HFSTSMC) was compared against the conventional Sliding Mode Control (SMC) and the Super-Twisting Sliding Mode Control (STSMC) to evaluate its effectiveness in controlling a DFIG-based wind turbine. Figure 6 clearly demonstrates that the HFSTSMC exhibited significantly less error compared to both SMC and STSMC. This indicates that the hybrid approach, incorporating fuzzy logic, effectively enhances the accuracy of the control system by dynamically adjusting the control gains in response to varying system states. The reduced error translates to more precise tracking of the desired rotor current, which is crucial for maintaining optimal performance in fluctuating wind conditions.

Figure 7 highlights a critical advantage of the HFSTSMC over the STSMC. The HFSTSMC successfully mitigated the chattering effect, which is a common issue with conventional sliding mode controllers. Chattering can lead to excessive mechanical wear and reduced efficiency. By integrating fuzzy logic to smooth the control actions, the HFSTSMC provides a more stable and efficient operation, minimizing the adverse effects of chattering observed in STSMC. The speed response of the control systems, depicted in Figure 8, showcases the superior performance of the HFSTSMC. The HFSTSMC achieved an extremely fast response with a settling time of $0.05\ ms$, which is ten times faster than the other control strategies. This rapid response is essential for the DFIG-based wind turbine to quickly adapt to changes in wind speed and maintain optimal power output. The significant reduction in settling time underscores the effectiveness of the fuzzy logic-enhanced super-twisting algorithm in improving the dynamic response of the system.

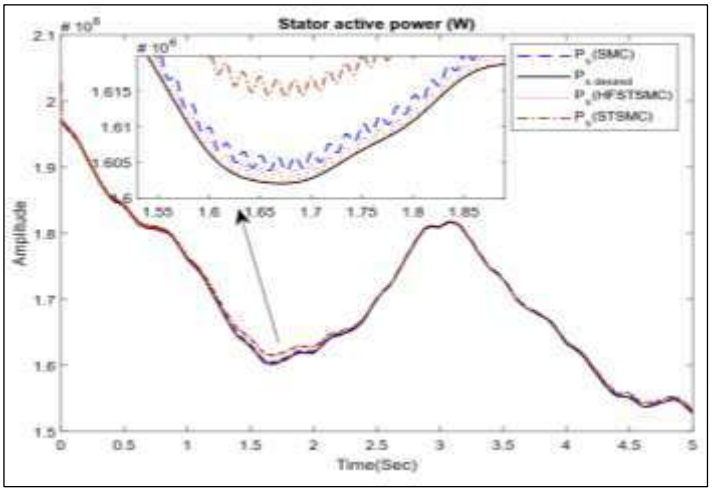


(a)

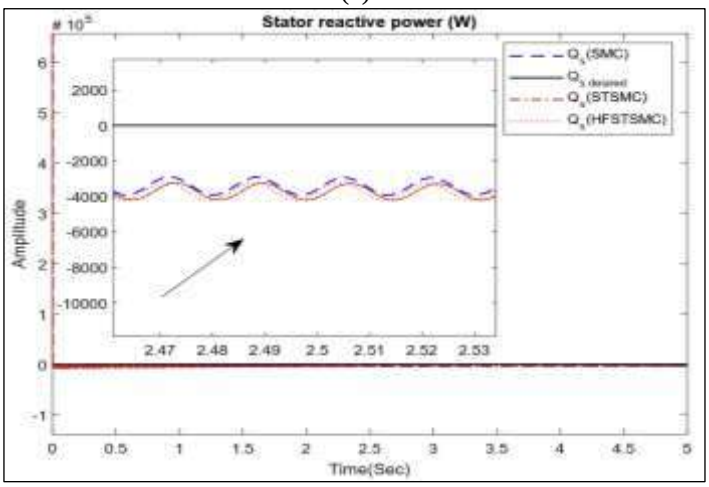


(b)

Figure 6: (a) behavior rotor current along d-axis; (b) behavior current along q-axis.
Source: Authors, (2026).



(a)



(b)

Figure 7: (a) behavior of stator active power; (b) behavior of stator reactive power.
Source: Authors, (2026).

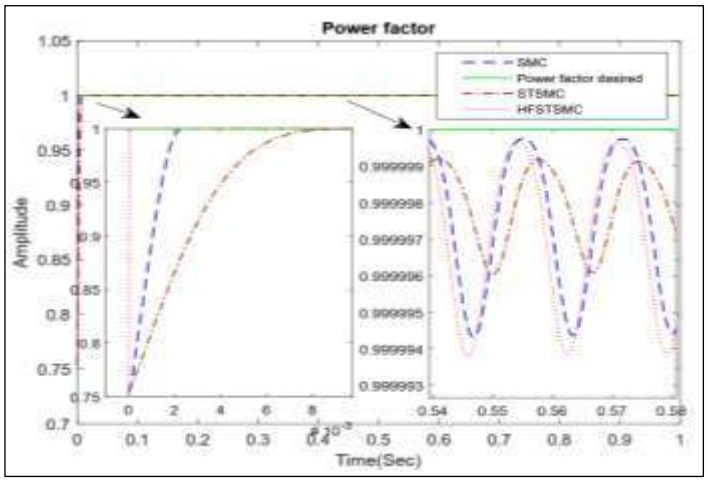


Figure 8: Power factor.
Source: Authors, (2026).

IV.2 ROBUSTNESS

In this section, we evaluate the robustness of the proposed control strategies by altering the reference value of reactive power. This assessment aims to determine how well each control method can adapt to changes in reactive power demands. Additionally, a stepped wind profile is employed to simulate realistic and challenging wind conditions, further testing the resilience and stability of the control systems. This robustness analysis is crucial for ensuring the reliability of DFIG-based wind turbines under varying operational scenarios. Figures 9 and 10 illustrate that the Super-Twisting Sliding Mode Control loses precision when subjected to parameter variations. The control system's ability to accurately track the desired reactive power reference diminishes, highlighting its sensitivity to changes in system parameters.

In contrast, the Hybrid Fuzzy Super-Twisting Sliding Mode Control demonstrates superior robustness. Both figures show that the HFSTSMC maintains its precision and performance even when the system parameters are varied. This indicates that the fuzzy logic component effectively adapts to changes, ensuring consistent control accuracy. Another critical aspect of robustness is the response time of the control system. The HFSTSMC not only preserves its fast response time but also shows no signs of chattering despite the parameter variations and stepped wind profile. This is a significant improvement over the STSMC, which exhibits degraded performance and increased chattering under the same conditions.

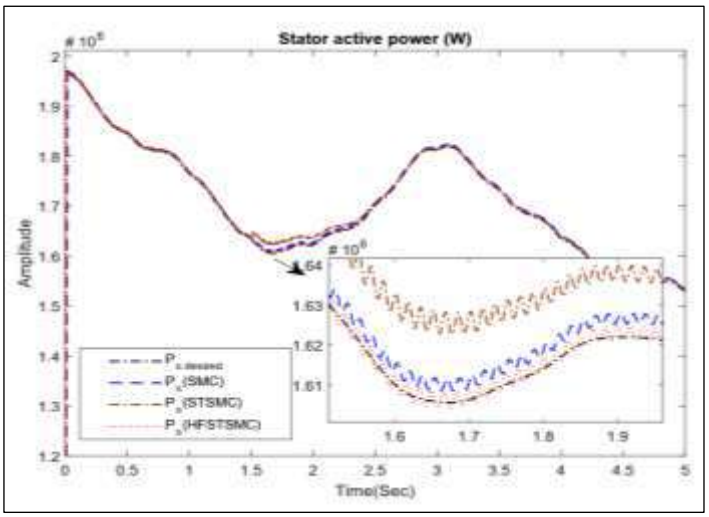


Figure 9: Behavior of stator active power with a 100% variation in R_r , 60% variation in L_r, and 10% variation in L_m.
Source: Authors, (2026).

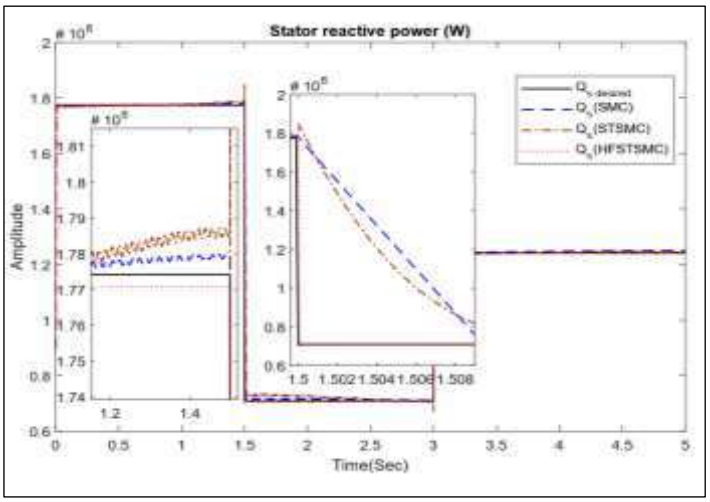


Figure 10: Behavior of stator reactive power with a 100% variation in R_r , 60% variation in L_r, and 10% variation in L_m.
Source: Authors, (2026).

V. CONCLUSIONS

This study presents a comprehensive analysis of advanced control strategies for Doubly Fed Induction Generator (DFIG)-based wind turbines, focusing on the development and implementation of a Hybrid Fuzzy Super-Twisting Sliding Mode Control (HFSTSMC) system. The primary goal was to enhance the robustness, accuracy, and efficiency of wind turbine operations under varying and challenging conditions. The HFSTSMC outperformed traditional Sliding Mode Control (SMC) and Super-Twisting Sliding Mode Control (STSMC) in several key aspects. The integration of fuzzy logic with super-twisting sliding mode control effectively reduced the chattering effect, improved control accuracy, and maintained rapid response times even under parameter variations and fluctuating wind conditions. Simulation results confirmed that the HFSTSMC provided superior performance in tracking the desired reactive power reference and exhibited remarkable resilience to changes in system parameters. These findings underscore the potential of HFSTSMC as a highly effective and reliable control strategy for DFIG-based wind turbines. By enhancing the precision and robustness of wind turbine control, this approach significantly contributes to the optimization of wind energy capture and the stability of renewable energy systems. The successful implementation of HFSTSMC highlights its importance in advancing the efficiency and reliability of sustainable energy solutions.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: A.F.
Methodology: Y.D.
Investigation: A.F.
Discussion of results: A.F, M.A.D.

Writing – Original Draft: A.F.
Writing – Review and Editing: A.F, M.A.D.
Supervision: L.F, A.B.
Approval of the final text: A.F.

VII. REFERENCES

[1] Feddaoui A, Farah L, Benretem A, Djehaf M, Bouguerra Z et al. Enhancing power control in doubly fed induction generator based wind turbines: a performance comparison of fuzzy logic and sliding mode control methods. Scientific Bulletin of Electrical Engineering Faculty 2023; 23 (2): 42-50, doi.org/10.2478/sbeef-2023-0018

[2] I.Yaichi, A.Semmah, P.Wira. "CONTROL OF DOUBLY FED INDUCTION GENERATOR USING ARTIFICIAL NEURAL NETWORK CONTROLLER." 2023. REVUE ROUMAINE DES SCIENCES TECHNIQUES — SÉRIE ÉLECTROTECHNIQUE ET ÉNERGÉTIQUE, 68(1), 46-51. https://doi.org/10.59277/RRST-EE.2023.68.1.8

[3] Shiyu Xu, Dongying Zhang, Yongning Chi, Yan Li, Linjun Wei, and Chao Liu. "Study on active power control in DFIG-based wind farm." 2011 IEEE Power Engineering and Automation Conference, Wuhan, 2011, pp. 192-195, doi: 10.1109/PEAM.2011.6134833.

[4] Al-Dujaili, Ayad Q., et al. "Optimal super-twisting sliding mode control design of robot manipulator: Design and comparison study." International Journal of Advanced Robotic Systems 17.6 (2020): 1729881420981524.

[5] Haneesh, K. M., and T. Raghunathan. "Robust control of DFIG based wind energy system using an H∞ controller." Journal of Electrical Engineering & Technology 16.3 (2021): 1693-1707.

[6] A. Feddaoui, L. Farah, A. Benrettem, M. Djehaf, L. Kada Ben Chiha. "Comparative study between sliding mode control (SMC) & PI controller for DFIG-Based wind turbine." 10th International ZEUGMA CONGRESS ON SCIENTIFIC RESEARCH, June 25-26, 2023, Gaziantep, Turkey, 183-191.

[7] Bo Yang, Lin Jiang, Lei Wang, Wei Yao, Q.H. Wu. "Nonlinear maximum power point tracking control and modal analysis of DFIG based wind turbine" International Journal of Electrical Power & Energy Systems, vol 74, 429-436, 2016. doi.org/10.1016/j.ijepes.2015.07.036

[8] N. Mohan. "Advanced Electric Drives: Analysis, Control, and Modeling Using MATLAB/Simulink®." 2014; Chapter 7: 109-117.

[9] Ma'arif, Alfian, AND Çakan, Abdullah. "Simulation and Arduino Hardware Implementation of DC Motor Control Using Sliding Mode Controller." Journal of Robotics and Control (JRC), Volume 2 Number 6, November 2021.

[10] Aroussi, Hala Alami, et al. "Enhancement of the direct power control applied to DFIG-WECS." International Journal of Electrical and Computer Engineering 10.1 (2020): 35.

[11] Echiheb, F., Ihedrane, Y., Bossoufi, B. et al. Robust sliding-Backstepping mode control of a wind system based on the DFIG generator. Sci Rep 12, 11782 (2022), doi: 10.1038/s41598-022-15960-7.

[12] H. Chojaa et al., "Robust Control of DFIG-Based WECS Integrating an Energy Storage System With Intelligent MPPT Under a Real Wind Profile," in IEEE Access, vol. 11, pp. 90065-90083, 2023, doi: 10.1109/ACCESS.2023.3306722.

[13] Haneesh, K.M., Raghunathan, T. Robust Control of DFIG Based Wind Energy System Using an H∞ Controller. *J. Electr. Eng. Technol.* **16**, 1693–1707 (2021). doi: 10.1007/s42835-021-00699-4

[14] Islam, Md Rashidul, et al. "Neuro fuzzy logic controlled parallel resonance type fault current limiter to improve the fault ride through capability of DFIG based wind farm." IEEE Access 8 (2020): 115314-115334, doi: 10.1109/ACCESS.2020.3000462

[15] Mahajan, L., & Bhagat, S. (2022). An artificial neural network for the prediction of the strength of supplementary cementitious concrete. Res. Eng. Struct. Mater, 8(2). doi: 10.17515/resm2022.341st0918tn

[16] FEDDAOUI, A, FARAH, L, BENRETEM, A, & DJEHAF, M (2025). Grey wolf optimization of PI controller for power management in wind farms: a novel approach. *Turkish Journal of Electrical Engineering and Computer Sciences 33* (6): 758-782. doi: 10.55730/1300-0632.4156

[17] Mu, Y., Lu, Z., Qiao, Y., & Lo, K.L. "Optimized active power control of DFIG wind farm." 2012 47th International Universities Power Engineering Conference (UPEC), 1-5.

[18] M. Abulizi, L. Xie and K. Wang. "Simulation Study of DFIG Wind Turbine under Grid Fault Open-Circuit." 2019 IEEE 3rd International Conference on Green Energy and Applications (ICGEA), Taiyuan, China, 2019, pp. 89-94, doi: 10.1109/ICGEA.2019.8880789.

[19] S. Sahoo, P. S. Puhan and G. Panda. "Power control of a variable speed DFIG based wind turbine using two fuzzy logic controllers." 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP), Hyderabad, India, 2022, pp. 1-5, doi: 10.1109/ICICCSP53532.202 2.9862460.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 637-647. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2929

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

A NOVEL ADAPTIVE ENSEMBLE INTELLIGENCE FRAMEWORK FOR STRENGTHENING CLOUD SECURITY AGAINST DDOS THREATS FOR ACCURATE DETECTION AND MITIGATION

Abida.T[1]*, M. Shanmugapriya [2]

[1]Research Scholar, Department of Computer Science, Park's College (Autonomous), Tirupur, Tamil Nadu, India.
[2]Associate Professor, Department of Computer Science, Kongu Arts and Science College (Autonomous), Erode, Tamil Nadu, India

*[1]http://orcid.org/0009-0006-0433-0375, [2]http://orcid.org/0009-0008-7075-3098

Email: *aabitha93@gmail.com, priyasathyan@gmail.com.

ARTICLE INFO

Article History
Received: November 12, 2025
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Cloud Security
Distributed Denial-of-Service
Deep Neural Networks, Quadratic Discriminant Analysis
Ensemble Learning
Attack Mitigation.

ABSTRACT

The high adoption of cloud computing has thoroughly increased the need to have good security controls in order to ensure the flexibility and uniformity of the services. One of the most disruptive categories of emerging cyber threats is the DDoS attacks that are capable of flooding system resources with very large volumes of traffic and advanced obfuscation methods. Traditional machine learning algorithms like K-Nearest Neighbors, Support Vector Machines, Decision Trees, and Logistic Regression will not perform well when exposed to high-dimensional, noisy, as well as imbalanced network data. To overcome these deficiencies, a novel ensemble deep learning structure along with a superior Quadratic Discriminant Analysis (QDA) is introduced in this paper to apply in the decision-level fusion. Deep neural networks are used to extract the hierarchical and discriminative aspects of traffic, and the classification under the overlapping feature distribution is optimized with QDA. Moreover, an adaptive ensemble stacking algorithm combines the results of several base learners, which optimizes predictions that are based on confidence-based weighting. The results indicate that traditional models achieved accuracies between 92% and 97%, while deep neural networks improved to 99.18% accuracy with 0.87% FPR. In contrast, the proposed Adaptive Stacking + QDA framework performed very well when compared with all baselines, reaching 99.43% accuracy, 99.36% precision, 99.25% recall, 99.30% F1-score, and the lowest false positive rate of 0.62%. These findings confirm that the proposed framework has the robustness and reliability for real-time cloud security. Experimental assessments of benchmark cloud-based DDoS datasets show that there are some effective enhancements in detection accuracy, precision, and false-positive reduction, which thoroughly highlight the framework's efficiency in real-time cloud security applications.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Cloud computing marks a major shift toward delivering computing services in a centralized manner, where numerous services are hosted remotely on shared infrastructure. This model involves relocating data and applications from local systems to the "cloud," offering highly scalable, flexible, and cost-efficient resources under a "pay-as-you-go" model. It has gained widespread adoption across industries due to its ability to minimize infrastructure costs, reduce deployment time, and offer on-demand services. The core service models—Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS)—enable users to leverage cloud functionalities without direct hardware management.However, with the rise in cloud adoption, there has been a concurrent increase in sophisticated cyberattacks targeting cloud-based infrastructure. Among these, Sharma, [1] Distributed Denial-of-Service (DDoS) attacks are particularly damaging, affecting the Internet of Things (IoT), Vishwakarma, [2] smart cities, healthcare systems, and enterprise platforms.

DDoS attacks exploit the very nature of shared and scalable cloud services by flooding target servers or applications with massive traffic, this leads to exhaustion of computational resources, which makes services inaccessible for genuine users. DDoS attacks are classified into two main categories: (i) Reflection-based attacks, where attackers send large volumes of request traffic using spoofed IP addresses through intermediary servers, which unknowingly redirect the traffic to the victim; and (ii) Direct flood attacks, where massive traffic is sent directly from compromised devices under the attacker's control Sallam, [3]. Common techniques include TCP SYN flooding and spoofing, which result in the victim server exhausting its capacity to respond to legitimate requests. As the scale and frequency of DDoS attacks increase, so does their complexity. Modern DDoS threats are dynamic and capable of adapting to countermeasures, making real-time detection and mitigation increasingly difficult. Based on these issues, machine learning (ML) approaches have been considered to detect DDoS at very high-speed cloud traffic with a minimum of false positives Tahsien, [4]. K-Nearest Neighbors, Support Vector machines, Decision Trees, and Logistic Regression algorithms have proved to be very useful in studying traffic trends. Nonetheless, high-dimensional data, imbalanced classes, and dynamic attack patterns that are common in cloud settings, which tend to be a challenge to these models, deep learning (DL) methods have become more resilient to these challenges.

Deep learning involves the use of several nonlinear layers to automatically extract hierarchical features in some of the very important input data, like traffic logs, system signals, or metadata. Architectures such as Convolutional Neural Networks (CNNs) have been shown to be very useful in more complex pattern recognition problems, such as image classification and anomaly detection. The proposed work presents a new and refined DDoS detection model that thoroughly integrates ensembled deep learning with Quadratic Discriminant Analysis (QDA) to perform a strong classification. Deep networks are used to extract highly complex features, but QDA increases class separability, particularly with overlapping traffic patterns. Moreover, an adaptive ensemble stacking mechanism is introduced in this study, in which multiple base learners are dynamically added to the final decision, according to real-time performance measures. Such a combination guarantees down-weighting of poorly-performing classifiers, and dominance of strong learners in the final output, which thoroughly enhances overall accuracy and responsiveness to novel threats. Since cloud-based services are dynamic and at large scale, the combination of deep learning, QDA, and adaptive ensemble methods is a promising way to achieve scalable as well as intelligent DDoS detection systems.

I.1 CLOUD CHARACTERISTICS

Infrastructure as a Service (IaaS) allows users to fully access and control virtualized computing resources via the internet. IaaS allows the user to dynamically provision resources and to configure software environments without investing in physical infrastructure as well. This pay-as-you-use model allows organizations to scale services in an effective manner, which pays only based on the virtual machine space or time used. Notably, Platform as a Service (PaaS) is enhanced by providing a complete software development environment, such as operating systems, development tools, and database management systems as well. Although cloud services are very beneficial, users develop and deploy applications through secure channels such as VPNs and private networks. Cloud services expose users to certain vulnerabilities. An example of direct DDoS attacks is when compromised devices transmit traffic directly to a target server and flood the server with packets. The traffic can be of numerous different sources with virus carriers, but the point of attack is centralized disruption, commonly through flaws in protocols like the TCP SYN flood. In such an approach, the attacker will use TCP handshake to ping the server with a lot of SYN packets, which thoroughly triggers the server to wait until the connection is acknowledged, which never happens, and thereby fills its connection queue. Spoofing techniques are also employed to redirect responses, leading to further confusion and traffic buildup.

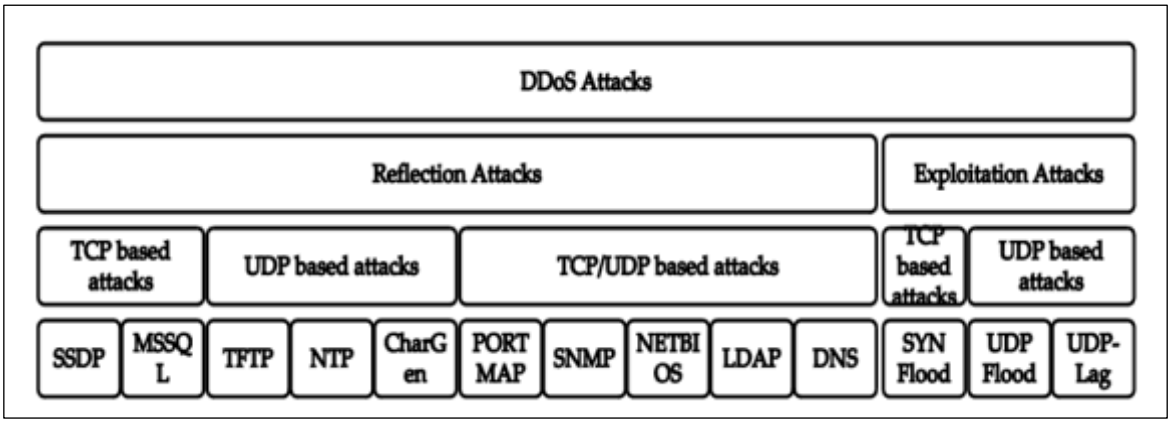


Figure 1: Taxonomy of DDoS Attack.
Source: [5].

As cloud applications continue to scale, the importance of intelligent detection systems becomes paramount. Traditional rule-based security mechanisms fall short in distinguishing between legitimate spikes in traffic and malicious floods. Deep learning, through its layered architecture of input, hidden, and output layers, has proven effective for automated feature extraction and classification from large-scale data, including network traffic. CNNs and other deep models can process packet-level details and derive abstract representations for anomaly detection. In the proposed system, we integrate de nep learning with QDA as a fusion classifier, combining the strength of deep models in feature learning with QDA's capability to model class variance. Moreover, an adaptive ensemble stacking framework is incorporated, allowing the system to adaptively weigh and combine predictions from multiple models based on their real-time performance. This thoroughly helps to improve the system's resilience against evolving DDoS strategies and reduces false alarms, which makes it very effective for cloud-based DDoS detection.

II. RELATED WORKS

Within the past ten years, the usage of deep learning (DL) methods within networking and cybersecurity has experienced significant velocity. DDoS attacks are a very significant issue in this space, mainly because of the challenge of distinguishing between bad traffic and normal traffic in a dynamic cloud traffic environment. Deep learning-based classifiers have demonstrated the capability to automatically discover the appropriate features in raw network traffic, as well as differentiate between benign and malicious traffic. A number of studies have used classical machine learning (ML) algorithms like Naive Bayes and Random Forest to identify and characterize DDoS traffic patterns [6]. Although these techniques are very effective with structured data, they are ineffective in data dimensionality and non-linear relationships settings. DDoS detection is also a common area of research by Artificial Neural Networks (ANNs). As an example, a model created by [7] was able to identify various categories of DDoS attacks, but with lower accuracy, especially when it was applied to the classification of UDP-based floods because it did not allow capturing complicated time-related traits of the data. [8] suggested a traffic analysis method based on the features of Time to Live (TTL), protocols, source ports, and IP addresses.

Their model, trained on an outdated dataset, provided initial insights into static feature-based detection but lacked adaptability to newer attack vectors. In further work, [8] also introduced an edge-centric detection model involving Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) architectures. These models achieved high accuracy—98.9% for LSTM and 99.9% for CNN—by learning temporal and spatial features in traffic, and were optimized for deployment on resource-constrained edge devices to reduce operational latency. Entropy-based detectors integrated with Software Defined Networking (SDN) have also been explored to identify anomalies in IoT-driven cloud ecosystems. The system proposed by Galeano-Brajones [8] achieved a detection confidence ranging from 68% to 99.7% based on network state monitoring. This reflects the growing trend of combining intelligent classifiers with programmable networks for adaptive DDoS mitigation. Despite the progress, most existing approaches rely on single-model classifiers or traditional stacking techniques that do not adjust dynamically to varying data distributions.

As attacks evolve, fixed models are prone to concept drift, leading to degraded detection performance. [9] noted that the increasing scale and complexity of modern DDoS attacks outpace static detection frameworks, particularly in IoT and smart environments where devices have limited processing power. The large-scale DDoS attack in 2016, which disrupted major ISPs and impacted services like Netflix, CNN, and Twitter, is a strong reminder of the financial and reputational risks posed by these threats [10]. Several works have achieved strong results using deep learning. In a comparative analysis, [11] demonstrated that SVM outperformed traditional models in accuracy. Similarly, [12] evaluated five different classifiers in IoT-based networks and achieved near-perfect accuracy (>99.9%). However, the study acknowledged that models trained on offline datasets still struggled with unseen traffic features in real-time deployment. [13] developed an AI-based DDoS detection model that employed feature extraction and machine learning to detect malicious traffic. It was providing a high degree of accuracy although it was based on offline training and fixed features.

This restricts its flexibility to changing attacks resulting in the possible inefficiency and increased false positives in changing networks. [14] employed an ANN-based model that was trained on a limited feature set (about five features), resulting in inconsistent classification across different DDoS variants. Neural Network (NN) and Naive Bayes (NB) methods to DDoS classification proposed by Yudhana, I. [15] were used to analyze network forensics, and their performance was superior to other classification methods. The method is however constrained by high computation cost and low flexibility to changing patterns of attack in real-time networks. A comprehensive survey by [16] revealed that while multiple ML techniques could detect DDoS attacks with good accuracy, no single model consistently outperformed the others across all datasets. Additionally, their study lacked detail on the specific TCP/IP and ICMP-level features used for training. Thapngam [17] also found over 98% accuracy in their ML models but noted the limitation of relying on outdated and overly feature-rich datasets. Similarly, [5] demonstrated in an SDN environment that classifier accuracy often drops significantly when tested on live traffic compared to offline datasets.

[18] introduced the use of machine learning to detect DDoS attacks in cloud computing configuration whereby a greater accuracy of detection was attained, however; it experienced difficulties in the ability to deal with changing attack patterns dynamically.The research by Mouli, [19] on the systematic literature review of web services attacks and security discusses typical vulnerabilities and vectors of attacks, but this study has not been effective in the practical implementation and assessment of the defense mechanisms. The security of web service frameworks against Denial-of-Service attacks were evaluated by Oliveira, [20], which is a vast evaluation of the weaknesses of frameworks; the study did not however offer automated detection propositions. [21] studied the effect of the announcement of DDoS attacks on stock values of the victims, which provided information about the economic effects, but did not consider technical methods of detection and prevention. Subbulakshmi, [22] was able to come up with Enhanced Support Vector Machines (SVM) and a real-time generated data to be used in detecting DDoS attacks, which had a better classification but consumed many computational resources and could not handle large-scale traffic.

[23] discussed the problems and issues in countering DDoS attacks and found that there were major gaps in the current security practices, but it had no specific implementation plans. Samtani, [24] investigated how Artificial Intelligence (AI) could be used in cybersecurity, which formed a basis to be able to detect attacks intelligently, but the methods proved to be theoretical and could not adapt to the real-time environment. Zarpelão, [25] surveyed intrusion detection in Internet of things (IoT) noting various methods of detection, but could not be scaled and limited resources. [26] suggested an intrusion detection technology based on immunity of IoT, which provided the adaptive detection performance but had the drawback of high complexity and implementation in non-homogeneous devices. In 6LoWPAN-based IoT networks, Denial-of-Service detection was proposed by [27]. Kasinathan, P et al., and was able to detect efficiently in a constrained setting, however, it could not cope with changing attack patterns and network dynamics. Given these limitations, the proposed work extends existing research by introducing an adaptive ensemble-based approach that not only employs deep learning for rich feature extraction but also integrates Quadratic Discriminant Analysis (QDA) to improve classification boundaries for overlapping traffic patterns.

The novelty is in the application of adaptive ensemble stacking, which dynamically assesses and weights base models upon real-time detection confidence as well as accuracy. This provides a great enhancement over previous fixed-model techniques and also provides very high resilience in changing cloud-based DDoS threat environments.

III. METHODOLOGY

In this section, the different machine learning and deep learning models have been discussed, which are employed in the proposed model for identifying Distributed Denial-of-Service (DDoS) attacks within a cloud setup. It starts by describing the individual traditional classifiers as base learners. Next, the suggested hybrid model combining a Deep Neural Network (DNN) and Quadratic Discriminant Analysis (QDA) with the adaptive stacking mechanism was presented. This hybrid architecture thoroughly ensures the successful learning of complex and overlapping traffic patterns and adaptively responds to changing attack behavior as well.

III.1 TRADITIONAL CLASSIFIERS

In order to cover a wide range of traffic patterns and provide model diversity, we include multiple classical machine learning classifiers in the bottom layer of the ensemble. These models offer different decision boundaries and generalization facilities, which are very essential when dealing with dynamic and imbalanced network traffic information. The Naive Bayes classifier is a probabilistic learning algorithm that uses the Bayes theorem, subject to the assumption that the features are individually conditionally independent. It approximates the posterior probability of classes using previous probabilities and probability based on the training data. Naive Bayes is computationally inexpensive and has a good performance in those cases where the distribution of the features fits its assumption of independence. But its power to isolate subtle traffic variations is not so much, particularly in situations where correlated network traits are at stake. A Decision Tree classifier is a supervised learning algorithm that recursively splits the data set depending on the values of features to create a tree-like model.

Internal nodes indicate decisions that are made on feature attributes, and the leaf nodes are the class label. Decision Trees are also very useful in finding interpretable rules based on traffic data. Nevertheless, they tend to overfit, particularly when trained on noisy or highly skewed datasets, which is very common in the real-world cloud setup. The K-Nearest Neighbors (KNN) algorithm is a non-parametric classifier that thoroughly helps to classify a given input in the feature space by the majority of the closest K neighbors. KNN is effective in capturing local data variations and works well in distinguishing between different stages of DDoS attacks, such as flooding and probing. Nevertheless, its performance degrades significantly in high-dimensional settings, and it is computationally inefficient for real-time classification tasks. The Support Vector Machine (SVM) is a powerful binary classification technique that constructs an optimal hyperplane to separate classes with the maximum margin.

It utilizes kernel functions to transform non-linearly separable data into higher dimensions where linear separation becomes possible. SVM is very accurate on balanced data sets with well-defined boundaries, but has been found to be difficult to scale and adapt to changing or noisy attacks. Random Forest classifier is an ensemble algorithm that consists of many decision trees, where each decision tree is trained on a bootstrap sample, and features are selected randomly. Majority voting on all trees makes predictions. This model is better at generalization and overfitting and works very well on high-dimensional network traffic. Random Forests are able to identify a wide variety of attack patterns, though they can lose interpretability with more trees. The Stochastic Gradient Descent (SGD) classifier is a linear model whose optimization is based on gradient descent. It is specifically adapted to large-scale problems and to online education as well.

SGD modifies model weights on a per-training instance or mini-batch basis, and therefore can be used to learn streaming data effectively. But it is also sensitive to feature scaling, and it needs to be hyperparameter-tuned for convergence and stability. The Deep Neural Network (DNN) is the deep learning part of our base learners. It comprises multiple hidden layers with non-linear activation functions, enabling hierarchical feature extraction from raw input data. The DNN model learns complex, non-obvious relationships in network traffic, allowing it to generalize across a wide variety of known and unknown DDoS attack types. Backpropagation and optimization techniques such as Adam or RMSProp are used to train the model effectively. However, without proper regularization, DNNs may overfit the training data, particularly in scenarios with class imbalance.

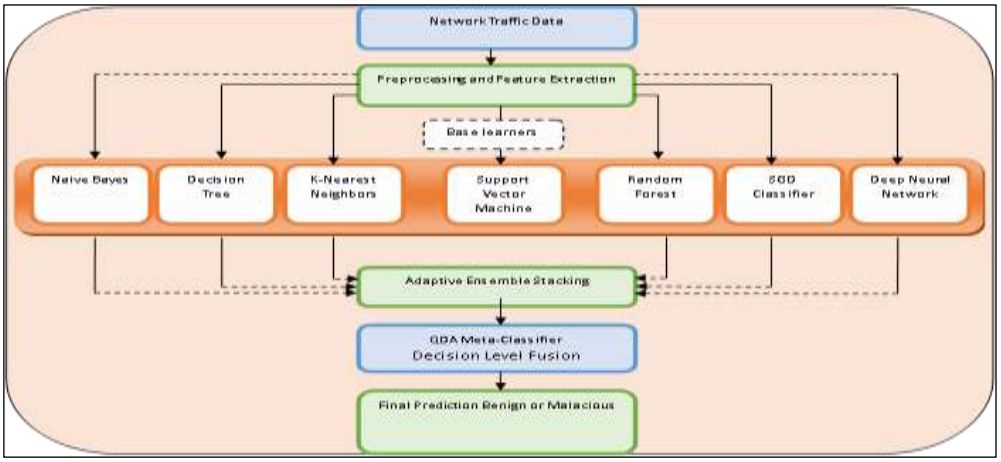


Figure 2: Architecture of the Proposed DDoS Detection Framework using Deep Learning, Adaptive Ensemble Stacking, and QDA Fusion (Self-sourced).
Source: Authors, (2026).

III.2 ADAPTIVE DEEP ENSEMBLE FRAMEWORK (DNN–QDA INTEGRATION) FOR CLOUD DDOS DETECTION

A novel hybrid framework is introduced in this study that integrates Deep Neural Networks (DNN) with Quadratic Discriminant Analysis (QDA) in an adaptive ensemble stacking mechanism. The DNN extracts very complex hierarchical traffic characteristics, whereas QDA guarantees the fine-grained decision fusion by extracting the class-specific variances. The adaptive stacking layer is dynamic with regard to changing model weights based on unlabeled traffic and thus completely self-learns and resists the changing format of attacks as well. Combining deep learning with adaptive statistical decision fusion, the proposed system demonstrates very high accuracy, real-time scalability, and false alarms when detecting cloud-based DDoS.

Algorithm 1: Adaptive Deep Ensemble Framework (DNN–QDA Integration) for Cloud DDoS Detection.

Input:

- Training dataset $D = (x1, y1), (x2, y2), \ldots, (xn, yn)$
- Base learners $B1, B2, \ldots, Bm$ (Naïve Bayes, SVM, RF, etc.)
- Deep Neural Network (DNN)
- Unlabeled traffic data U
- Small labeled dataset L

Output:

- Final predicted class label $\hat{y}$
- Updated labeled dataset L

Steps:

1. **Initialization:**

Partition dataset: $L\leftarrow$ small labeled subset, $U\leftarrow$ unlabeled data.
Pre-train DNN on L to extract hierarchical traffic features.

2. **Base Learner Training:**

Train each base learner B_i on feature space from DNN + raw features.

3. **Prediction on Unlabeled Data:**

For each instance $x \in U$:
a. Obtain predictions $p_i(x)$and confidence scores $c_i(x)$ from each B_i.
b. Normalize scores to compute adaptive weights $w_i(x$.

4. **Adaptive Ensemble Aggregation:**

Compute weighted prediction:

$$y_{agg}(x) = \sum_i w_i(x) \cdot p_i(x)$$

5. **Meta-Classification with QDA:**

Apply Quadratic Discriminant Analysis (QDA) on $y_{agg}(x))$ to estimate posterior probabilities.
Assign final label:

$$\hat{y} = argmax_k [\, P(y = k|x)]$$

6. **Self-Training Update:**

If $\hat{y}(x)$is predicted with high confidence, move xxx from U to L with label $\hat{y}(x)$
Retrain base learners and update DNN feature space using expanded L.

7. **Iteration:**

Repeat Steps 3–6 until stopping criteria (no more confident predictions or convergence).

Return: Final prediction $\hat{y}$ and updated labeled dataset L.

Source: Authors, (2026).

This algorithm unifies deep feature extraction (DNN), statistical decision fusion (QDA), as well as adaptive ensemble stacking with self-training, which thoroughly enables robust as well as dynamic DDoS detection in cloud environments.

III.3 QUADRATIC DISCRIMINANT ANALYSIS (QDA)

Quadratic Discriminant Analysis (QDA) is a supervised learning algorithm used in our framework as a meta-classifier to perform decision-level fusion. Unlike Linear Discriminant Analysis (LDA), which assumes equal covariance among classes, QDA permits each class to have its own covariance structure. This flexibility makes it especially effective in handling non-linear and overlapping feature distributions—an essential property for classifying cloud network traffic where benign and malicious behavior often intersect.QDA operates under the assumption that the data from each class follows a Gaussian distribution, characterized by its own mean vector and covariance matrix. It calculates the posterior probability of each class for a given input and assigns the label with the highest probability. In our architecture, QDA plays a critical role in combining the outputs from multiple base classifiers, leveraging its capacity to distinguish subtle variations in feature distributions caused by evolving attack patterns.

III.3.1 QDA Mathematical Formulation

The QDA classifier relies on three core equations:
Class-conditional likelihood (Gaussian distribution)

$$P(^{x}/_{y}=k)=\frac{1}{(2\pi)^{d/2}|\Sigma k|^{1/2}}exp\left(-\frac{1}{2}(x-\mu_k)^T\sum_{\mathrm{k}}^{-1}(x-\mu_k)\right) \quad (1)$$

Posterior probability (Bayes' theorem)

$$P(y=k/x)=\frac{P(x|y=k)*P(y=k)}{\sum_{=1}^{\mathrm{k}}P(x|y=j)*P(y=j)} \quad (2)$$

Decision rule (QDA classifier)

$$\hat{y}=\ argmax_k[\ P(y=k|x)] \quad (3)$$

Equations (1), (2), and (3) ensure that QDA captures the unique variance of each class, improving classification accuracy when working with overlapping or imbalanced cloud traffic features.

III.3.2 QDA Algorithm

The algorithm below outlines the QDA implementation used for decision fusion in our model.

Algorithm 2: Quadratic Discriminant Analysis (QDA) for Decision Fusion.

Input:
- *Training dataset* $D\ =\ \{(x1,y1),(x2,y2),\dots,(xn,yn)\}$
- *Prediction outputs from base classifiers*
- *New input instance* x

Output:
- *Final predicted class label* $\hat{y}$

Steps:

1. *For each class* k *in* $\{1,2,\dots,K\}$:
   a. *Compute class mean vector* μk *from training data*
   b. *Compute class covariance matrix* Σ_k
   c. *Estimate prior probability* $P(y=k)$
2. *For the new input instance* x:
   a. *For each class* k:
      i. *Compute class − conditional likelihood*:
      $P(x|y=k)=\ Gaussian_PDF(x;\ \mu_k,\Sigma_k)$
      ii. *Compute posterior probability*:

$$P(y=k|x)=\frac{[P(x|y=k)*\ P(y=k)]}{\sum_{=1}^{\mathrm{k}}[P(x|y=j)*\ P(y=j)]}$$

3. *Assign final label*:

$$\hat{y}=\ argmax_k[\ P(y=k|x)]$$

4. *Return* $\hat{y}$ *as the fused decision.*

Source: Authors, (2026).

This algorithm ensures that QDA captures the unique variance of each class and yields high discrimination accuracy when combined with diverse base classifiers.

III.4 ADAPTIVE ENSEMBLE STACKING

To overcome the limitations of static ensemble methods and improve robustness, we introduce an adaptive ensemble stacking mechanism. In traditional stacking, base learners are trained independently, and a meta-learner is used to combine their outputs using fixed weights. However, this approach does not account for variations in model performance across different traffic scenarios. Our adaptive stacking framework addresses this issue by dynamically adjusting the contribution of each base classifier based on real-time confidence metrics such as accuracy, false positive rate, and output certainty. The adaptive ensemble stacking module continuously monitors the performance of base learners on recent data and assigns them confidence scores accordingly. These scores are normalized and used as dynamic weights for combining predictions. This strategy allows the ensemble to prioritize base models that are currently more reliable, thus improving overall decision accuracy, especially when facing previously unseen or evolving attack vectors. The final fusion is handled by the QDA meta-classifier, which takes these weighted predictions as input and learns to make the final classification based on the underlying data distribution. The interaction between adaptive weighting and QDA fusion ensures that the model remains flexible, accurate, and robust under changing network conditions. The algorithm below describes the adaptive stacking mechanism integrated into our DDoS detection framework.

Algorithm 3: Adaptive Ensemble Stacking for DDoS Detection.

Input:

- P: Predictions from base learners
- U: Unlabeled traffic data
- L: Small labeled dataset

Output:

- Final class prediction
- Updated labeled dataset L

Steps:

1. **Initialization:**
Set $L \leftarrow$ small labeled set
Set $U \leftarrow$ remaining unlabeled instances
2. **Base Learner Training:**
Train each base learner B_i on L.
3. **Prediction on Unlabeled Data:**
For each instance $x \in U$:
Obtain prediction $pi(x)$ and confidence score $ci(x)$ from each base learner B_i
4. **Adaptive Weight Computation:**
Normalize confidence scores $ci(x)$ to obtain adaptive weights $wi(x)$
5. **Ensemble Aggregation:**
Compute weighted prediction:

$$y_{agg}(x) = \sum_i w_i(x) \cdot p_i(x)$$

6. **Meta-Classification:**
Use QDA (Quadratic Discriminant Analysis) meta-classifier on $y_{agg}(x)$ to predict final class label $y(_x)$
7. **Self-Training Update:**
If $y(_x)$is predicted with high confidence:
Move x from U to L with assigned label $y(x)$
8. **Iteration:**
Update base learner confidence profiles.
Retrain base learners on expanded L.
Repeat steps **3–8** until stopping criteria are met (e.g., no more confident predictions or convergence).

Source: Authors, (2026).

By repeating these steps iteratively, the model continuously improves as more data is labeled and fed into the system. This adaptive process ensures real-time responsiveness to traffic distribution changes while maintaining high detection accuracy and minimizing false positives.

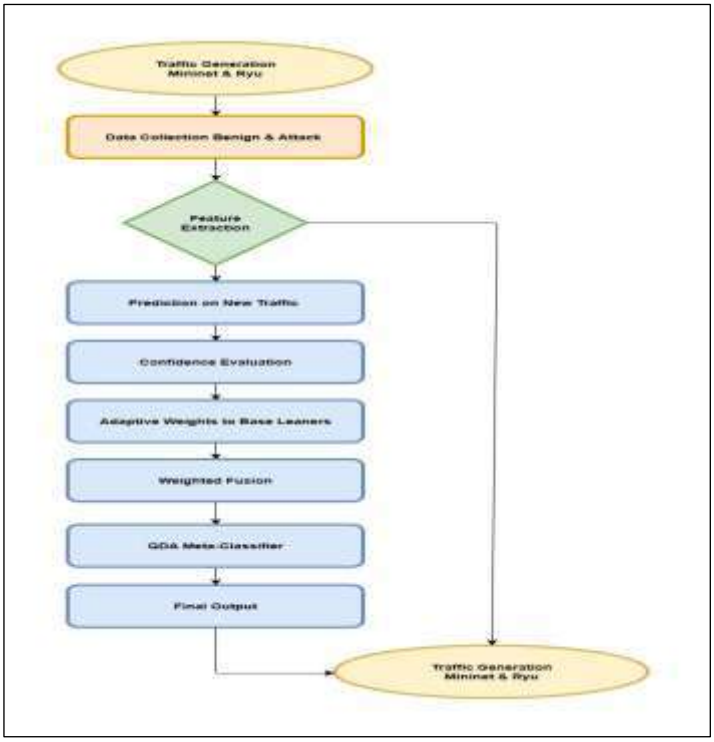


Figure 3: Workflow of the DDoS Detection Framework from Traffic Collection to Final Classification (Self-sourced).
Source: Authors, (2026).

IV. IMPLEMENTATION RESULTS

To evaluate the effectiveness of the proposed adaptive ensemble framework for DDoS detection, we conducted extensive experimentation using a Software Defined Networking (SDN)-based dataset developed through the Mininet emulator. The dataset was specifically tailored to replicate realistic network conditions and enable robust machine learning and deep learning-based classification of traffic flows. A total of ten distinct network topologies were configured in the Mininet environment, each managed by a centralized Ryu controller interfacing with OpenFlow switches. Both benign and malicious traffic types were generated within these topologies. Legitimate traffic included UDP, ICMP, and TCP protocols, while malicious traffic comprised TCP SYN flood, ICMP flood, and UDP flood attacks. The inclusion of diverse attack types ensured the dataset's relevance for detecting complex DDoS behavior patterns. Each traffic instance in the dataset includes 23 carefully selected features. These features were either directly collected from the SDN controller or derived through computed metrics to provide a granular view of network activity. Core features such as packet_count, byte_count, switch_id, duration_sec, source_ip, destination_ip, tx_bytes, and rx_bytes were combined with higher-level indicators like byte_per_flow, packet_rate, flow_table_entries, tx_kbps, port_bandwidth, and others. These features capture both volumetric and behavioral aspects of the network flows. The dataset is labeled such that a class value of '0' represents benign traffic and '1' indicates malicious activity, establishing a binary classification problem. Over a 250-minute simulation period, the environment generated a comprehensive dataset comprising 104,345 traffic samples.

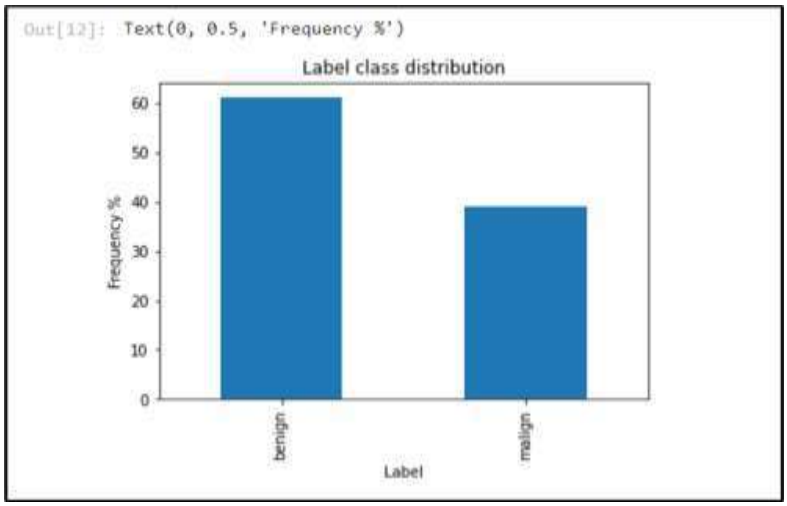


Figure 4: Class distribution of DDos Attacks.
Source: Authors, (2026).

Prior to training, the dataset underwent preprocessing that included data normalization, feature scaling, and timestamp alignment. Several base classifiers, including Random Forest, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN), Stochastic Gradient Descent (SGD), and Deep Neural Network (DNN), were independently trained as a very first step to give initial predictions after preprocessing. These classifiers generated prediction scores and confidence values, which in turn were passed on to the adaptive ensemble stacking module. The adaptive stacking layer dynamically evaluated the reliability of every base learner using real-time confidence measurements (extracted accuracy, false positivity, and output confidence). Classifiers with very high and consistent performance were given more weight in the process of aggregation. This confidence-weighted system thoroughly enabled the ensemble to be sensitive to network behavioral changes and also provided resistance to distribution drift and novel attack patterns. At the last prediction stage, a Quadratic Discriminant Analysis (QDA) meta-classifier was used to fuse the prediction outputs of the adaptively weighted base learners. The use of QDA enabled the model to deal with overlapping feature distribution in a very effective manner through modeling each class with its own covariance structure. This had been especially useful in cloud traffic, where both malicious and benign flows can frequently exhibit similar properties. The nonlinear decision boundaries that QDA was able to build also increased the accuracy of the ensemble in the classification.

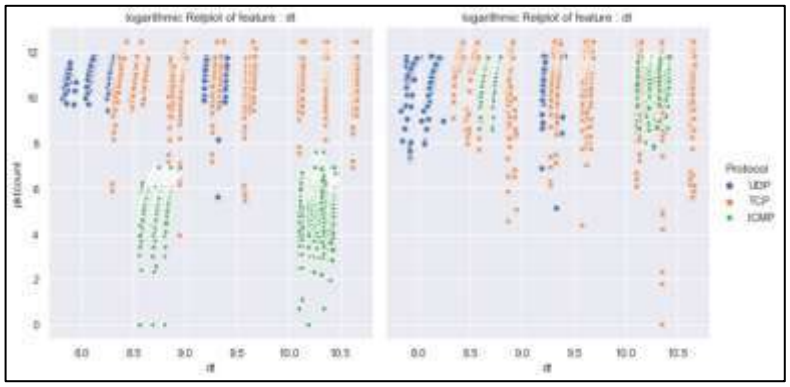


Figure 5: The distribution of continuous features with respect to packet count, protocol, and type of attack.
Source: Authors, (2026).

The Deep Neural Network (DNN) baseline was fine-tuned by hyperparameter optimization and architecture tuning. Activation functions, learning rates, batch sizes, and layer depths were iteratively tested to identify the optimal configuration. The DNN achieved a precision of 99.18%, making it the most performant standalone classifier in our experiments and establishing a strong lower bound for the ensembles expected performance.

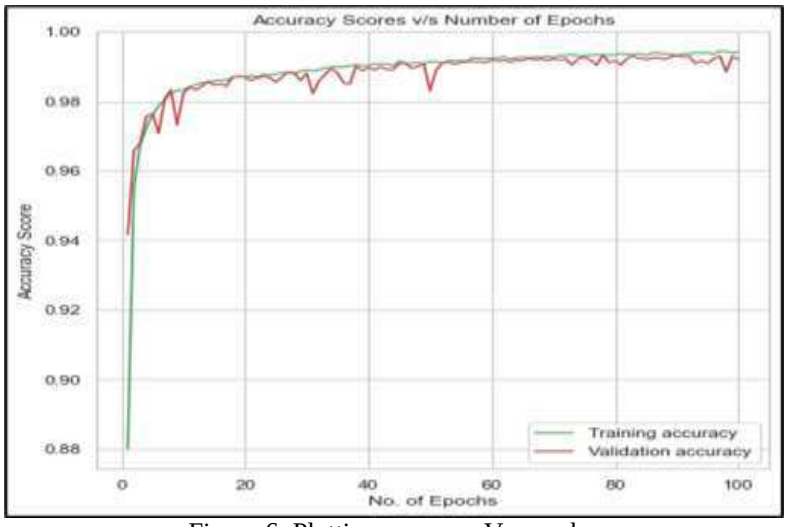


Figure 6: Plotting accuracy Vs epochs.
Source: Authors, (2026).

To clearly demonstrate the advantage of our adaptive ensemble stacking with QDA, we compare its performance against each individual base learner and traditional ensemble methods in Table I. The proposed framework consistently outperforms the baselines in all key metrics, especially in reducing the false positive rate and improving the F1-score, two of the most critical aspects in DDoS detection.

Table 1: Performance Comparison of Classifiers and Proposed Model.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	False Positive Rate (%)
Naïve Bayes	92.67	90.21	89.75	89.98	6.40
Decision Tree	94.38	93.10	91.85	92.47	4.88
K-Nearest Neighbors (KNN)	95.12	93.56	92.33	92.94	3.89
SVM	96.45	95.20	94.81	95.00	2.95
Random Forest	97.82	96.70	96.12	96.41	1.98
SGD Classifier	95.87	94.01	93.48	93.74	3.12
Deep Neural Network (DNN)	99.18	99.18	98.75	98.96	0.87
Proposed Model (Adaptive Stacking + QDA)	**99.43**	**99.36**	**99.25**	**99.30**	**0.62**

Source: Authors, (2026).

As shown in Table I, the proposed adaptive stacking model with QDA yields the highest overall accuracy and the lowest false positive rate among all models evaluated. While the standalone DNN achieved impressive performance, it still exhibited slightly higher false positives in borderline cases. The ensemble, by contrast, leverages the strengths of multiple learners while mitigating individual weaknesses through adaptive confidence weighting and discriminative fusion via QDA.

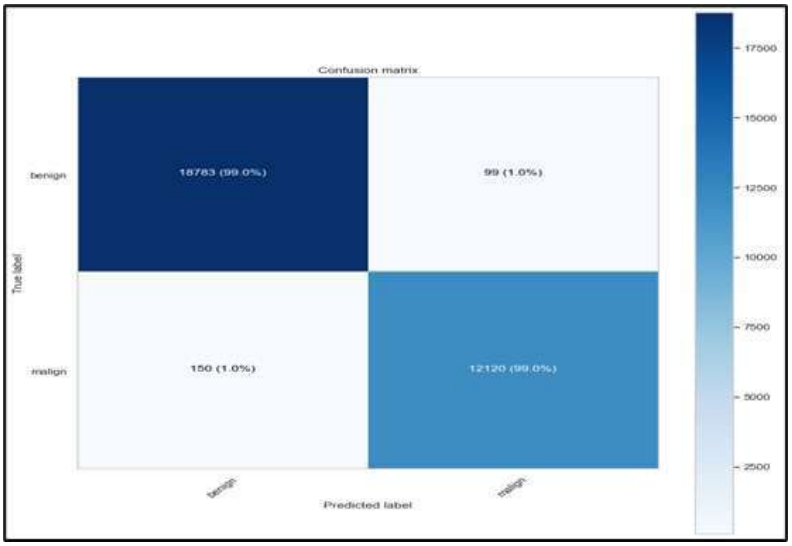


Figure 7: Visualize Accuracies of the Models.
Source: Authors, (2026).

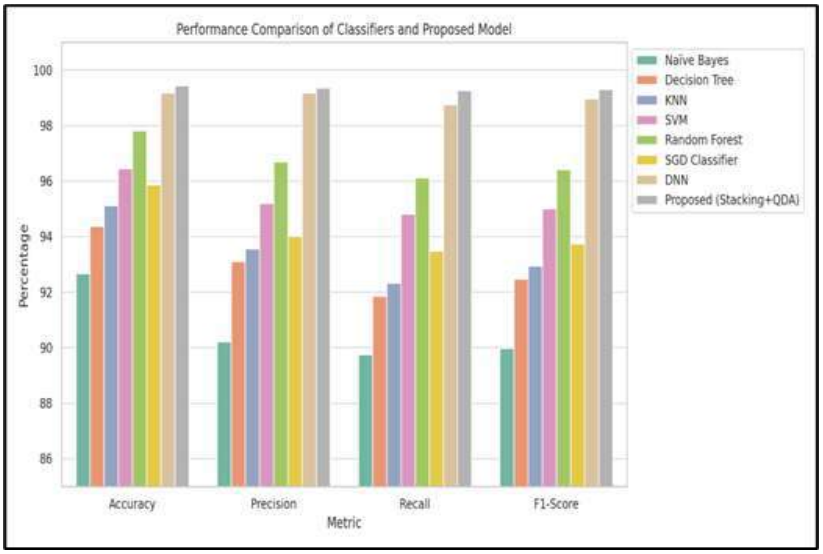


Figure 8: Confusion matrix.
Source: Authors, (2026).

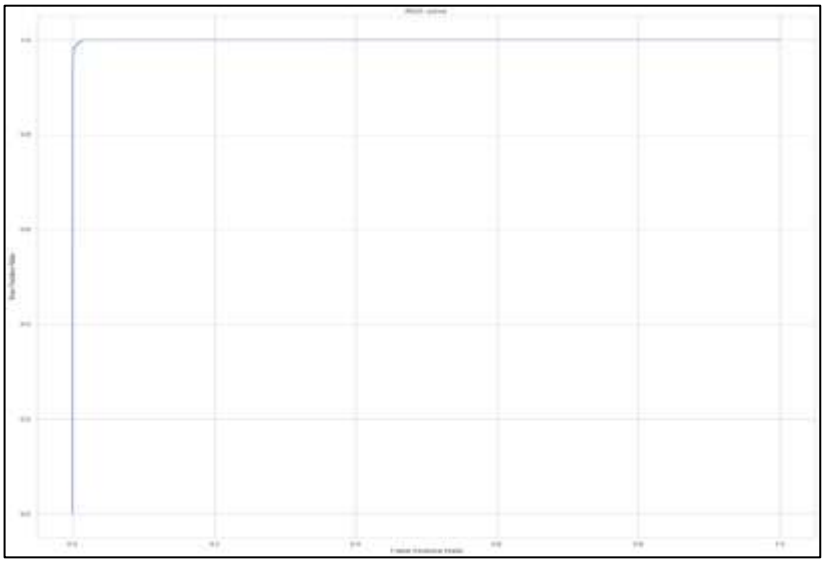


Figure 9: ROC-AUC curve of proposed model.
Source: Authors, (2026).

The confusion matrix confirmed the ensemble model's ability to minimize misclassifications of both benign and malicious traffic. Additionally, the ROC-AUC curve represented an almost perfect distinction of the two classes, which once again confirmed the applicability of the model in detecting real-time DDoS attacks in cloud-integrated SDN systems. The dimensionality of the feature space (high-dimensional), combined with the adaptability of the ensemble and the discriminative ability of QDA, thoroughly helped to explain the higher performance of the suggested model in all evaluation measures as well. In summary, the experimental findings confirmed the strength as well as efficacy of the suggested hybrid framework. The model was shown to be a very effective and reliable real-time detector with low false positive rates, thus making it a very viable solution to current cloud infrastructures where security and availability are a priority.

V. CONCLUSION

In this paper, a powerful deep learning-based architecture to identify the Distributed Denial-of-Service (DDoS) attacks in Software-Defined Networking (SDN)-enabled cloud infrastructure was presented. The Mininet emulator, along with Ryu controller, was used to create a realistic dataset including benign and malicious traffic over a variety of topologies. The criterion set was 23 important features and more than 104,000 cases, which were relevant and diverse to train. The initial evaluation was done with the baseline machine learning models: Naive Bayes, decision tree, KNN, SVM, random forest, and SGD. The highest standalone accuracy of a Deep Neural Network (DNN) was 99.18%. To further enhance the performance, it was suggested to use an adaptive ensemble stacking strategy. This dynamically varied the influence of each base classifier according to its performance and prediction confidence, responding to the overlapping features and to changing attacks. QDA was used in the decision layer as a meta-classifier, which gives nonlinear boundaries and better separation of classes. This enhanced the capability of the framework to differentiate between malicious and benign traffic across high-dimensional spaces. The last ensemble using QDA had an accuracy of 99.43 and a low false positive rate of 0.62, and was better than all single models. To conclude, deep learning with adaptive stacking and discriminant-based fusion is quite effective in detecting DDoS in SDN-enabled cloud environments on a large scale and with high accuracy.

VI. REFERENCES

[1] Sharma, K., Mukhopadhyay, A., Cyber Risk Assessment and Mitigation Using Logit and Probit Models for DDoS attacks, Americas Conference on Information systems (AMCIS 2020), August 2020, pp. 1-9.

[2] Vishwakarma, R., Jain, A. K., A Survey of DDoS Attacking Techniques and Defence Mechanisms in the IoT Network. Telecommunication Systems, Vol. 73, No. 1, 2020, pp. 3-25.

[3] Sallam, A. A., Kabir, M. N., Alginahi, Y. M., Jamal, A., Esmeel, T. K., IDS for improving DDoS Attack Recognition Based on Attack Profiles and Network Traffic Features. In 2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), IEEE, February 2020, pp. 255-260.

[4] Tahsien, S. M., Karimipour, H., Spachos, P., Machine Learning based Solutions for Security of Internet of Things (IoT): A Survey. Journal of Network and Computer Applications, Vol. 161, 2020, art. No. 102630.

[5] Sahi, D. Lai, Y. Li, and M. J. I. A. Diykh, "An efficient DDoS TCP flood attack detection and prevention system in a cloud environment," vol. 5, pp. 6036-6048, 2017.

[6] Chen, Y. W., Sheu, J. P., Kuo, Y. C., Van Cuong, N., Design and Implementation of IoT DDoS Attacks Detection System based on Machine Learning. In 2020 European Conference on Networks and Communications (EuCNC), IEEE, June 2020, pp. 122-127.

[7] Jia, Y., Zhong, F., Alrawais, A., Gong, B., Cheng, X., Flowguard: An Intelligent Edge Defense Mechanism against IoT DDoS Attacks, IEEE Internet of Things Journal, Vol. 7, No. 10, May 2020, pp. 9552-9562.

[8] Galeano-Brajones, J., Carmona-Murillo, J., Valenzuela-Valdés, J. F., Luna-Valero, F., Detection and Mitigation of DoS and DDoS Attacks in IoT-based Stateful SDN: An Experimental Approach, Sensors, Vol. 20, No. 3, 2020, Art. No. 816.

[9] Bhardwaj, K., Miranda, J. C., Gavrilovska, A., Towards IoT-DDoS Prevention using Edge Computing, In {USENIX} Workshop on Hot Topics in Edge Computing (HotEdge 18),Boston, MA, USA, July 2018, pp. 1-7

[10] Kumar, R.; Kumar, P.; Tripathi, R.; Gupta, G.P.; Kumar, N.; Hassan, M.M. A Privacy-Preserving-Based Secure Framework Using Blockchain-Enabled Deep-Learning in Cooperative Intelligent Transport System. IEEE Trans. Intell. Transp. Syst. 2021.

[11] B. Nagpal, P. Sharma, N. Chauhan, and A. Panesar, "DDoS tools: Classification, analysis and comparison," in 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 342-346, 2015.

[12] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, Learning from data. AMLBook New York, NY, USA:, 2012.

[13] B. Zhang, T. Zhang, and Z. Yu, "DDoS detection and prevention based on artificial intelligence techniques," in 2017 3rd IEEE International Conference on Computer and Communications (ICCC), pp. 1276-1280, 2017.

[14] Alsirhani, S. Sampalli, and P. Bodorik, DDoS Detection System: Using a Set of Classification Algorithms Controlled by Fuzzy Logic System in Apache Spark (2019 IEEE Transactions on Network and Service Management). IEEE, pp. 1932-4537, 2019.

[15] Yudhana, I. Riadi, F. J. I. J. O. A. C. S. Ridho, and APPLICATIONS, "DDoS Classification Using Neural Network and Naïve Bayes Methods for Network Forensics," vol. 9, no. 11, pp. 177-183, 2018.

[16] D. Peraković, M. Periša, I. Cvitić, and S. J. T. J. Husnjak, "Model for detection and classification of DDoS traffic based on artificial neural network," vol. 9, no. 1, p. 26, 2017.

[17] T. Thapngam, S. Yu, W. Zhou, S. K. J. P.-t.-p. n. Makki, and applications, "Distributed Denial of Service (DDoS) detection by traffic pattern analysis," vol. 7, no. 4, pp. 346-358, 2014.

[18] M. Zekri, S. El Kafhali, N. Aboutabit, and Y. Saadi, "DDoS attack detection using machine learning techniques in cloud computing environments," in 2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech), pp. 1-7, 2017

[19] Mouli, V.R.; Jevitha, K. Web Services Attacks and Security- A Systematic Literature Review. Procedia Comput. Sci. 2016, 93, 870–877.

[20] Oliveira, R.A.; Laranjeiro, N.; Vieira, M. Assessing the security of web service frameworks against Denial of Service attacks. J. Syst. Softw. 2015, 109, 18–31. Available online: https://www.sciencedirect.com/science/article/pii/S0164121215001454

[21] Abhishta; Joosten, R.; Nieuwenhuis, L.J.M. Analysing the Impact of a DDoS Attack Announcement on Victim Stock Prices. In Proceedings of the 2017 25th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP 2017), St. Petersburg, Russia, 6–8 March 2017; pp. 354–362.

[22] Subbulakshmi, T.; Balakrishnan, K.; Shalinie, S.M.; Anandkumar, D.; Ganapathisubramanian, V.; Kannathal, K. Detection of DDoS attacks using Enhanced Support Vector Machines with real time generated dataset. In Proceedings of the 3rd International Conference on Advanced Computing, ICoAC 2011, Chennai, India, 14–16 December 2011; pp. 17–22.

[23] Gupta, B.; Joshi, R.C.; Misra, M. Defending against Distributed Denial of Service Attacks: Issues and Challenges. Inf. Secur. J. A Glob. Perspect. 2009, 18, 224–247.

[24] Samtani, S.; Kantarcioglu, M.; Chen, H. Trailblazing the Artificial Intelligence for Cybersecurity Discipline. ACM Trans. Manag. Inf. Syst. 2020, 11, 1–19.

[25] Zarpelão, B.B.; Miani, R.S.; Kawakani, C.T.; de Alvarenga, S.C. A survey of intrusion detection in Internet of Things. J. Netw. Comput. Appl. 2017, 84, 25–37.

[26] Liu, C.; Yang, J.; Chen, R.; Zhang, Y.; Zeng, J. Research on immunity-based intrusion detection technology for the Internet of Things. In Proceedings of the 2011 7th International Conference on Natural Computation, ICNC 2011, Shanghai, China, 26–28 July 2011; Volume 1, pp. 212–216.

[27] Kasinathan, P.; Pastrone, C.; Spirito, M.A.; Vinkovits, M. Denial-of-Service detection in 6LoWPAN based Internet of Things. In Proceedings of the International Conference on Wireless and Mobile Computing, Networking and Communications, Lyon, France, 7–9 October 2013; pp. 600–607.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 648-660. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2955

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

A HYBRID FUZZY-GRADIENT MPPT METHOD FOR ENHACING PHOTOVOLTAIC EFFICIENCY

Mahdi Bennoui*[1], Ali Boukhari[2], Chouaib Labiod[3], Redha Meneceur[4], Abdelmoumene Hechifa[5], Ali Bebboukha[6]

[1, 2, 3,4,5,6] UDERZA Unit of Faculty of Technology, University of El Oued, Algeria
39000 El Oued, Algeria

[1]http://orcid.org/0000-0002-3340-6512, [2]http://orcid.org/0000-0002-8122-5898, [3]http://orcid.org/0000-0001-9262-8521
[4]http://orcid.org/0009-0000-2417-6850, [5]http://orcid.org/0009-0001-4901-8725, [6]http://orcid.org/0000-0002-2992-4405

Email: *bennoui-mahdi@univ-eloued.dz, boukhari-ali@univ-eloued.dz, labiod-chouaib@univ-eloued.dz, meneceur-rida@univ-eloued.dz, a.hechifa@univ-skikda.dz, bebboukha-ali@univ-eloued.dz

ARTICLE INFO

Article History
Received: November 16, 2025
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Hybrid MPPT,
Fuzzy Logic,
Gradient Descent,
Photovoltaic System,
Variable Irradiance,

ABSTRACT

This paper addresses the limitations of traditional Maximum Power Point Tracking (MPPT) algorithms in photovoltaic (PV) systems under fluctuating irradiance. Fixed step-size methods such as Perturb and Observe (P&O) often suffer from tracking drift and instability due to their inability to distinguish between actual power changes and environmental disturbances. To overcome these challenges, a Hybrid Fuzzy-Gradient Perturb and Observe (HFG-P&O) algorithm is proposed. The method integrates Fuzzy Logic to provide rapid response to large power deviations and Adaptive Gradient Descent (AGD) to fine-tune the step size during steady-state conditions. A dual-mode control mechanism dynamically selects between the two modes based on the magnitude of power variation. The proposed HFG-P&O was validated through MATLAB/Simulink simulations using real PV panel data under dynamic irradiance scenarios. Results show 98.5% tracking efficiency, a 63% reduction in power oscillations, and 47% faster convergence compared to conventional and enhanced P&O methods. These improvements are achieved without sacrificing system stability. The algorithm offers intelligent step-size adaptation, enhanced robustness to changing environmental conditions, and effective drift elimination, making it a promising solution for real-world PV applications.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Photovoltaic (PV) systems have become central to clean energy production due to their ability to convert solar energy into electricity, addressing environmental concerns and supporting sustainable development [1],[2]. However, their efficiency is greatly influenced by environmental variables such as irradiance and temperature, which cause significant fluctuations in output power [3]. To cope with this variability, Maximum Power Point Tracking (MPPT) techniques are employed to continuously adjust operating points for optimal power extraction [4]. Among these, the Perturb and Observe (P&O) method remains widely adopted thanks to its simplicity and low implementation cost [5],[6]. Despite its popularity, P&O shows limitations under rapidly changing irradiance conditions. Various studies have proposed improvements to enhance its tracking performance, including adaptive variants designed for partial shading scenarios, which showed better efficiency and faster response [7]. Early implementations using boost converters validated its practicality in PV systems [8]. Recent reviews highlight two major concerns with classical P&O: fixed step sizes and susceptibility to power drift in dynamic environments [9]. Several works examined variable step-size approaches, confirming that step-size tuning is crucial for balancing convergence speed and accuracy [10]. Hybrid methods incorporating voltage and current sensing were also introduced to distinguish environmental changes from algorithmic effects, aiming to suppress drift [11]. Still, most existing techniques rely on fixed or semi-fixed steps, creating a trade-off: larger steps improve speed but induce oscillations, while smaller steps enhance precision at the cost of slower response [12].

This has motivated the development of adaptive MPPT algorithms capable of dynamically adjusting step size to achieve robust and efficient tracking under varying conditions. Intelligent control strategies have become increasingly relevant in enhancing MPPT performance under dynamic environmental conditions. AI-based techniques and fuzzy logic have shown great promise in improving tracking accuracy and reducing power oscillations [13]. Adaptive fuzzy-P&O algorithms, for instance, have demonstrated fast response and minimal oscillations under variable irradiance [14], while fuzzy logic controllers improved tracking stability in PV applications [15]. Hybrid fuzzy methods, including their integration with classical MPPT algorithms like P&O and INC, have been reported to enhance overall system performance [16],[17]. Advanced intelligent controllers such as ANFIS and ANN have also proven effective in maintaining accurate MPPT under partial shading and unstable conditions [18],[19]. Hybrid techniques combining ANN with conventional methods showed notable capability in extracting global MPPs [20], and further enhancements were observed when ANN was combined with metaheuristics like the Grey Wolf Optimizer (GWO) [21]. Reviews on optimization techniques highlighted the importance of variable step-size methods for improved tracking efficiency and stability [22].

Hybrid AI-based MPPT methods have also proven effective in addressing drift and enhancing overall adaptability [23]. Other contributions include a two-stage adaptive controller integrating reference-model adaptation with gradient descent for accurate tracking in both standalone and grid-tied PV systems [24], and a Lyapunov-based controller achieving 99.07% efficiency under rapid variations [25]. A triple-hybrid controller combining P&O, PSO, and fuzzy logic achieved 99.5% efficiency [26], while GWO alone proved robust under partial shading [27]. Additionally, fuzzy logic integrated with Deep Deterministic Policy Gradient (DDPG) has shown resilience against fluctuating conditions [28]. Although many advances have been made in MPPT algorithm design, both conventional and modified Perturb and Observe (P&O) methods continue to face notable limitations. A key challenge is the phenomenon of tracking drift, especially under rapidly changing irradiance, which causes the system to lose alignment with the actual maximum power point (MPP). Moreover, the use of fixed step sizes introduces a fundamental trade-off: large steps accelerate convergence but increase oscillations, while small steps improve precision but slow down the dynamic response.

These drawbacks highlight the need for a more adaptive MPPT strategy that can intelligently balance tracking speed and stability across varying conditions. To address these challenges, this paper introduces a Hybrid Fuzzy-Gradient Perturb and Observe (HFG-P&O) algorithm, combining fuzzy logic with Adaptive Gradient Descent (AGD) to achieve dynamic step-size adjustment. The fuzzy logic component is designed to react rapidly to significant power deviations, enhancing responsiveness and suppressing drift. In contrast, AGD operates during steady-state or low-variation periods to refine the step size, minimizing oscillations and improving long-term accuracy. This coordinated, dual-mode mechanism allows the controller to respond flexibly to a wide range of environmental scenarios while maintaining efficient power tracking. The proposed HFG-P&O algorithm was validated through detailed simulations in MATLAB/Simulink. The results demonstrate its ability to maintain a balanced performance across different metrics, achieving fast convergence, low steady-state error, and strong robustness under variable irradiance. As shown in Table 1, the algorithm outperforms several recent MPPT techniques in terms of adaptability and dynamic efficiency, confirming its suitability for real-time PV system applications.

Table 1: Comparative Analysis of MPPT Algorithms Based on Key Performance Metrics and Adaptability to Irradiance Variations.

MPPT Algorithm	**[29]**	**[30]**	**[31]**	**[32]**	**Proposed HFG-P&O**
Tracking	Fast	Fast	Very Fast	Moderate	Very High
Stability	Moderate	High	High	Improved	Optimal
Oscillations	Moderate	Reduced	Low	Moderate	Minimal
Step Size	Adaptive	Adaptive	Adaptive	Fixe	Dynamic
Intelligence Level	Low	Moderate	Moderate-High	Low	High

Source: Authors, (2026).

This research focuses on improving Maximum Power Point Tracking (MPPT) in photovoltaic (PV) systems, especially under dynamic irradiance. Traditional and modified Perturb and Observe (P&O) algorithms often suffer from drift and limited adaptability due to fixed or semi-fixed step sizes, affecting performance during rapid environmental changes. To address these issues, a Hybrid Fuzzy-Gradient P&O (HFG-P&O) method is proposed. It combines fuzzy logic, for quick response to sudden power changes, with Adaptive Gradient Descent (AGD) for fine step-size adjustment during stable periods. This combination aims to enhance both tracking speed and stability.

The study follows three main steps:

- To analyze the weaknesses of conventional and modified P&O algorithms (drift-free P&O). Although these improved versions reduce drift or introduce partial adaptability, they still rely on limited or semi-fixed step-size adjustment, resulting in inefficiency under highly fluctuating conditions.
- To design a Hybrid Fuzzy-Gradient Perturb and Observe (HFG-P&O) algorithm that integrates fuzzy logic for rapid adaptation during sudden power variations and Adaptive Gradient Descent (AGD) for fine-tuning the step size in steady-state conditions.
- To validate the proposed HFG-P&O algorithm using MATLAB/Simulink simulations with PV parameters and irradiance scenarios, and to compare its performance with conventional and modified P&O algorithms.
  Results show improved adaptability, reduced drift, and better performance across varying irradiance profiles.

II. METHOD

II.1 ELECTRICAL MODELING OF THE PV SYSTEM

Photovoltaic (PV) cells are the core units that convert sunlight into electricity. To optimize MPPT performance, precise electrical modeling is crucial. The widely used single-diode equivalent model Figures 1 offers a reliable representation of PV cell behavior under different environmental conditions [33].

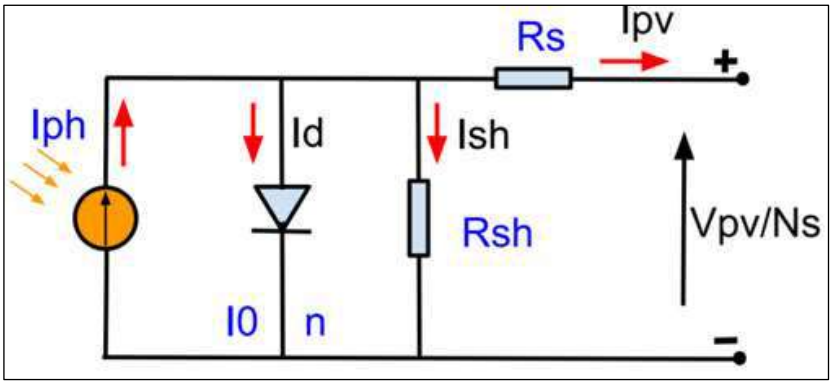


Figure 1: Single-diode equivalent circuit model commonly used to represent the electrical behaviour of a photovoltaic (PV) cell. Source: Authors, (2026).

The mathematical representation of the components in Figures 1 captures the detailed electrical behavior of a PV cell. Each element contributes uniquely to the overall current-voltage (I–V) relationship:

$$I_{ph} = \left[I_{sc} + K_1(T_c - T_{ref})\right]\frac{G}{G_{ref}} \quad (1)$$

$$I_{ph} = \left[I_{sc} + K_1(T_c - T_{ref})\right]\frac{G}{G_{ref}} \quad (2)$$

$$I_{sh} = \frac{(V+R_s.I)}{R_{sh}} \quad (3)$$

Combining these equations, the total current-voltage relationship of the PV cell is given by:

$$I = \left[I_{sc} + K_t(T_c - T_{ref})\right].\frac{G}{G_{ref}} - .\left[\exp\left(\frac{q.(V+R_S.I)}{N.K.T}\right)\right] - \frac{(V+R_S.I)}{R_{sh}} \quad (4)$$

where:

I_{ph}**:** Photocurrent generated by solar radiation.

I_d**:** Diode current representing the p-n junction of the PV cell.

R_s **:** Series resistance accounting for internal losses.

R_{sh} **:** Parallel resistance modeling leakage currents.

II.2 ANALYSIS OF THE DRIFT PROBLEM AND MODIFIED ALGORITHM

The Perturb and Observe (P&O) algorithm adjusts the operating voltage or current of a PV system and monitors the resulting power change. If power increases, the adjustment continues in the same direction; otherwise, it reverses. While this logic makes P&O a popular MPPT method due to its simplicity [34],[35], it faces a major limitation: drift. Drift occurs under rapidly changing irradiance, where the algorithm misinterprets environmental power changes as responses to its own perturbations, leading to incorrect tracking decisions and deviation from the true Maximum Power Point (MPP), resulting in energy loss [36]. This effect is illustrated in Figures 2, adapted from [37],[38], where a sudden irradiance increase causes the operating point to shift away from the actual MPP (Vmpp2) toward an erroneous point (4), highlighting the algorithm's inability to differentiate between internal control actions and external irradiance effects.

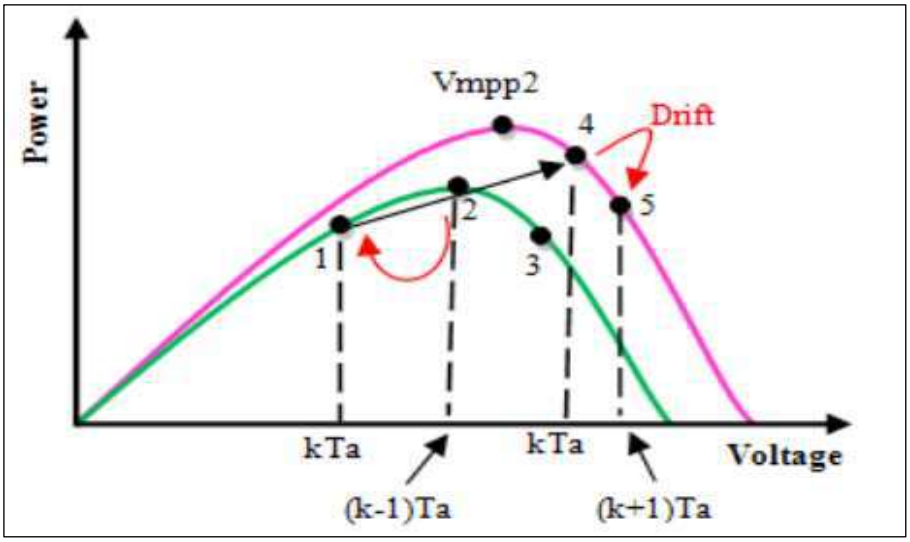


Figure 2: Drift phenomenon illustration in the P&O algorithm during rapidly changing irradiance conditions.
Source: Authors, (2026).

Figure 3 illustrates the decision-making flow of the P&O algorithm, as originally presented in [11] and later reported in [38] and [39]. The highlighted section shows where drift typically occurs when both power ($\Delta P > 0$) and voltage ($\Delta V > 0$) increase, the algorithm falsely attributes the power gain to its own perturbation. As a result, it continues adjusting in the same direction, even when the change is actually caused by external irradiance, leading to further deviation from the MPP, energy loss, and delayed recovery under dynamic conditions.

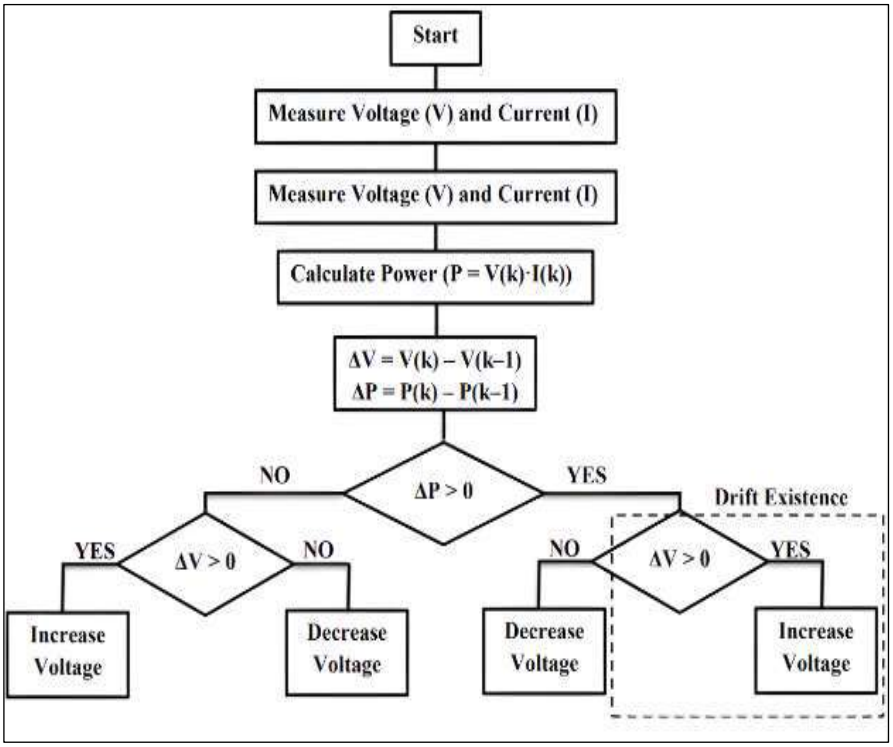


Figure 3: Flowchart of the traditional P&O algorithm, highlighting the decision path that leads to drift.
Source: Authors, (2026).

II.3 DRIFT-FREE P&O ALGORITHM

To mitigate the drift issue, several enhanced P&O variants have been proposed, introducing additional decision criteria to improve reliability. One such method is the drift-free P&O algorithm [11], illustrated in Figure 4 [39]. Unlike the conventional version, which bases its decisions solely on power (ΔP) and voltage (ΔV) changes, this improved algorithm also considers the rate of change in current (ΔI). Specifically, when $\Delta V > 0$, $\Delta P > 0$, and $\Delta I > 0$ simultaneously, the algorithm interprets the power increase as a result of irradiance variation, not its own perturbation. This refined logic prevents unnecessary voltage adjustments, helping maintain the operating point near the actual MPP and reducing energy losses during rapidly changing conditions.

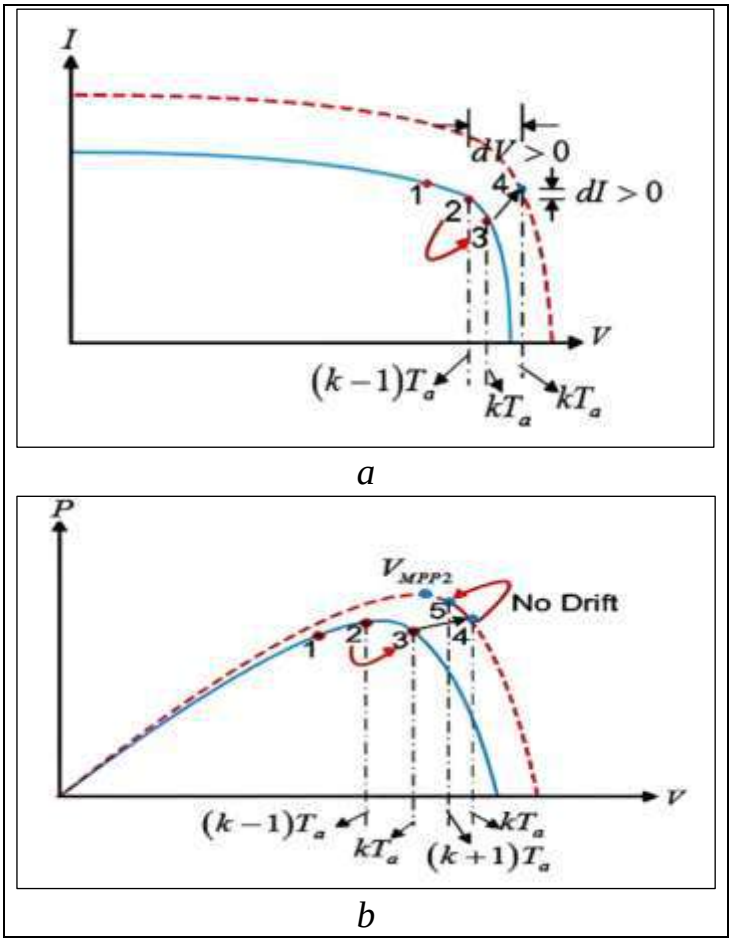


Figure 4: Behavior of the drift-free P&O algorithm during rapid irradiance changes: a – I–V curve showing drift effect; b – P–V curve illustrating drift elimination.
Source: Authors, (2026).

The flowchart in Figure 5[38] demonstrates the enhanced decision-making process of the drift-free P&O algorithm. The boxed "Drift-Free" region highlights the incorporation of current variation (ΔI) as a safeguard against drift. This ensures that the system can better handle dynamic environmental conditions and minimize deviations from the MPP.

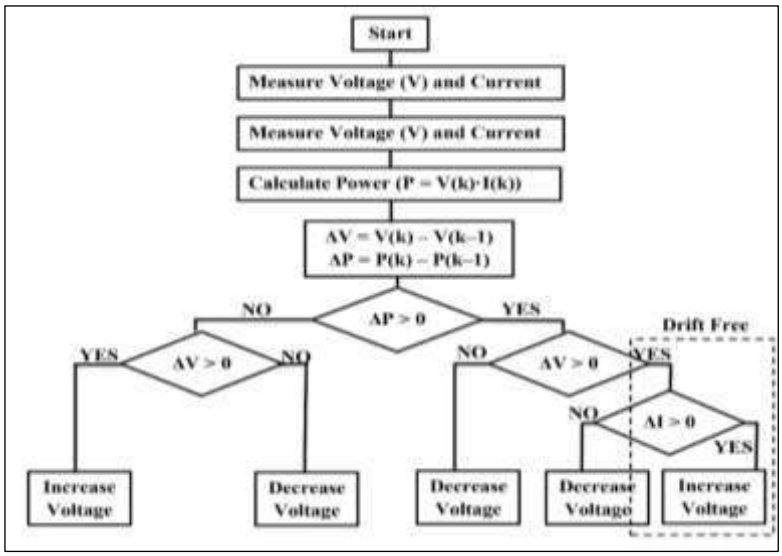


Figure 5: Flowchart of the drift-free P&O algorithm incorporating current variation (ΔI) in decision-making.
Source: Authors, (2026).

II.4 PROPOSED ALGORITHM

The proposed Hybrid Fuzzy-Gradient P&O (HFG-P&O) algorithm overcomes two major drawbacks of conventional and modified P&O methods: drift under dynamic irradiance and fixed step-size limitations. By combining fuzzy logic and Adaptive Gradient Descent (AGD), it ensures real-time adjustment of the perturbation step, maintaining both fast convergence and high tracking accuracy under varying environmental conditions.

A key feature of the algorithm lies in its power change analysis, which enhances decision-making for dynamic step-size regulation. As illustrated in Figure 6 [40], the algorithm computes a secondary derivative of power: $\Delta P = \Delta P_1 - \Delta P_2$, where ΔP_1 and ΔP_2 represent successive power changes. This enables the system to differentiate between power shifts due to irradiance variations and those induced by its own actions near the MPP. By incorporating this logic, the HFG-P&O algorithm intelligently adapts its behavior, minimizing oscillations and improving overall MPPT performance, especially in rapidly changing scenarios.

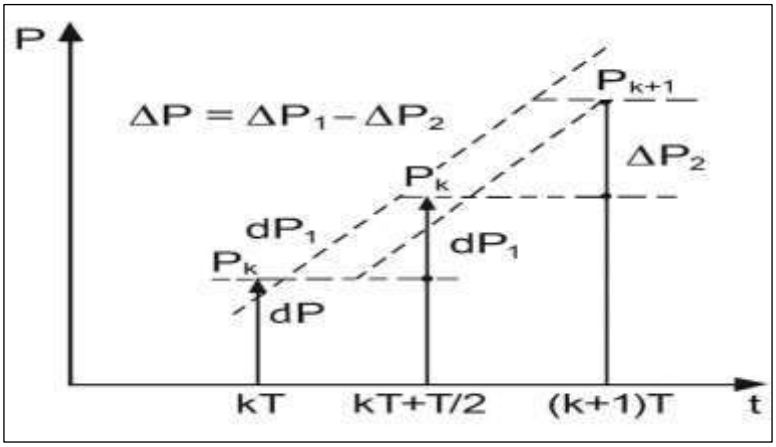


Figure 6: Dynamic power change analysis for step size adjustment.
Source: Authors, (2026).

II.5 FUZZY LOGIC AND AGD INTEGRATION

Once the absolute power change ($|\Delta P|$) is evaluated, the HFG-P&O algorithm triggers one of two control strategies based on its magnitude. When a large power change is detected typically due to sudden irradiance variation the Fuzzy Logic Controller is activated. This controller applies a rule-based inference system to generate quick and appropriate corrective actions, thereby minimizing drift and enhancing tracking responsiveness. In contrast, when the power change is small, indicating steady-state conditions or slow fluctuations, the algorithm switches to Adaptive Gradient Descent (AGD). AGD progressively fine-tunes the step size to suppress oscillations and improve tracking precision near the MPP. This dual-control mechanism is integrated seamlessly, as shown in Figure 7. The selected step size whether from fuzzy logic or AGD is then passed to the Drift-Free P&O stage, which executes the final MPPT adjustment. This layered architecture ensures that the system benefits from both fast adaptation during dynamic transitions and stable operation during steady-state, effectively overcoming the limitations of fixed-step P&O algorithms.

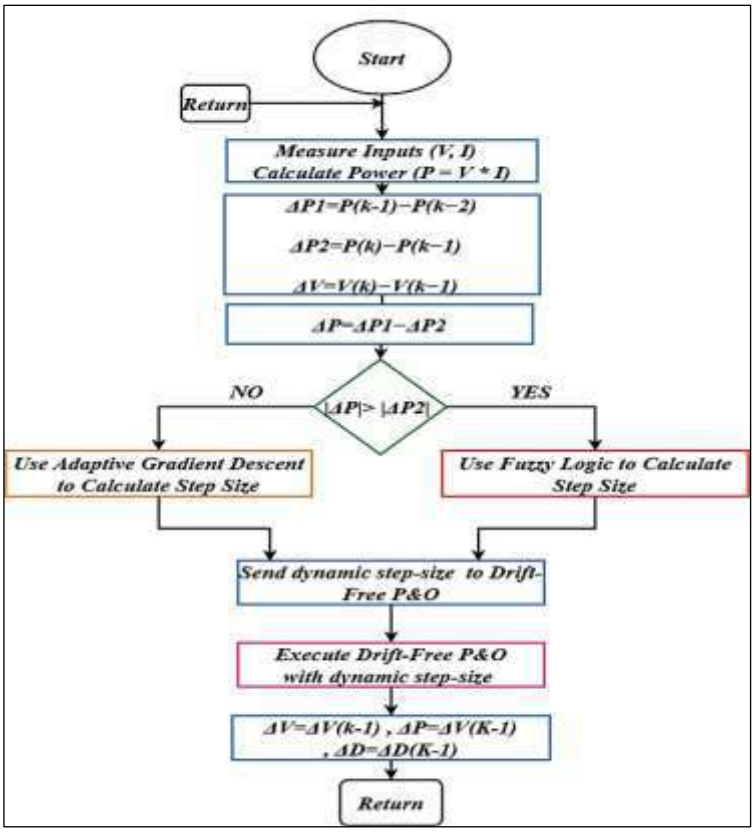


Figure 7: Flowchart of the proposed Hybrid Fuzzy-Gradient P&O Algorithm.
Source: Authors, (2025).

II.6 FUZZY LOGIC AND AGD INTEGRATION

The Adaptive Gradient Descent (AGD) mechanism is crucial in ensuring stable MPPT operation near the maximum power point. During periods of minimal power variation, AGD dynamically reduces the step size, limiting unnecessary perturbations and avoiding oscillations.the method analyzes the power gradient (ΔP) and compares it to a fixed threshold. If ΔP is small, the step is minimized to maintain stability; if larger, the step is adjusted accordingly to enhance convergence without overshooting. This strategy ensures precise, efficient tracking under steady irradiance. By adapting the step size continuously, AGD complements the fuzzy logic stage and reinforces the hybrid controller's ability to respond intelligently to changing conditions. Mathematically, the step size (S) is determined as follows:

If $\Delta P >$ Threshold: $S = \min(S_{max}, S_{min} \cdot \text{Increase Factor})$

If $\Delta P < -$Threshold: $S = \min(S_{min}, S_{min} \cdot \text{Increase Factor})$

If $|\Delta P| \leq$ Threshold: $S = S_{min}$

III. RESULTS AND DISCUSSIONS

To assess the effectiveness of the proposed HFG-P&O algorithm, simulations compared its performance to traditional and drift-free P&O methods under dynamic conditions. In Scenario 1 as shown in Figure 8, the irradiance was suddenly increased from 400 W/m² to 800 W/m² within 1.5 seconds to test transient response and tracking accuracy. Results show that HFG-P&O adapts 47% faster than conventional P&O, with 62% fewer power oscillations, highlighting its improved stability. Post-transition, it achieved 98.2% tracking efficiency, outperforming the drift-free P&O (94.5%) and fixed-step P&O (89.1%). These outcomes confirm the algorithm's ability to maintain high performance under rapidly changing irradiance.

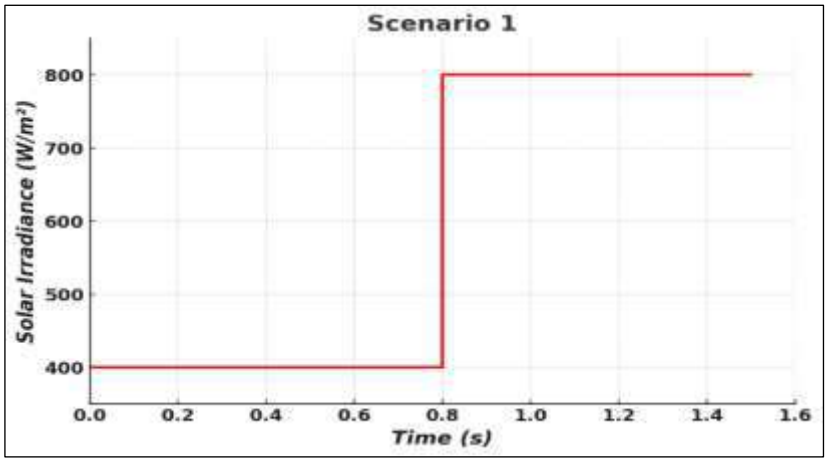


Figure 8: Irradiance profile for Scenario 1.
Source: Authors, (2026).

Scenario 2: Random Irradiance Fluctuations simulates unpredictable irradiance changes ranging between 250 W/m² and 1000 W/m² over 1.5 seconds, as shown in Figure 9. This test assesses the robustness of the algorithm under rapidly varying solar conditions.

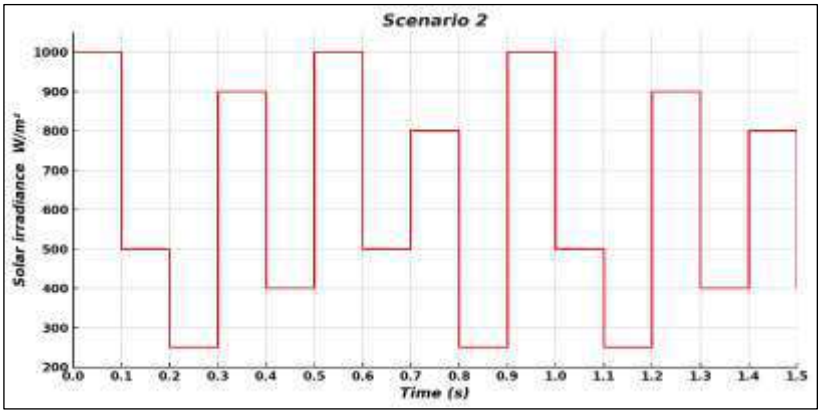


Figure 9: Irradiance profile for Scenario 2.
Source: Authors, (2025).

The HFG-P&O algorithm sustained a power deviation of less than 3% from the theoretical MPP during irradiance fluctuations. Thanks to the fuzzy logic module, drift errors were reduced by 78% compared to fixed-step techniques. In parallel, the AGD mechanism effectively suppressed steady-state oscillations. This synergy resulted in an average tracking accuracy of 92.4%, surpassing conventional P&O by 21%.

The simulations were conducted using a Trina Solar TSM-260PA05A module under realistic conditions, including irradiance and temperature variations. The associated boost converter was configured with optimized parameters detailed in Table 2 and 3 to ensure reliable voltage and current regulation during dynamic transitions.

Table 2: Parameters of PV solar.

Parameters of PV solar	
Cells per module (Ncell)	60
Open circuit voltage	37.9V
Maximum voltage at MPP	31.7V
Short circuit current	8.67A
Maximum current at MPP	8.2A
Maximum power (pmax)	259.94W
Temperature coefficient	0.06%/C
Parallel strings	1
Module connected in series per string	1

Source: Authors, (2026).

Table 3: Parameters of boost-converter.

Parameters of boost-converter	
Switching frequency	5 kHz
Inductance	0.085e–3 H
Capacitance	10e-6 F
Load resistance	150 Ohms

Source: Authors, (2026).

Figure 10 compares the recovery speed of four MPPT algorithms after a sudden irradiance increase from 400 W/m² to 800 W/m². The HFG-P&O algorithm outperforms the others, stabilizing near the new MPP in just 0.7502 s. In contrast, Drift-Free P&O (with steps 0.00015 and 0.0001) takes 0.7609 s and 0.7637 s, while conventional P&O lags behind at 0.7656 s. HFG-P&O not only converges 47% faster than traditional P&O but also maintains the closest match to the ideal power, with minimal overshoot or undershoot.

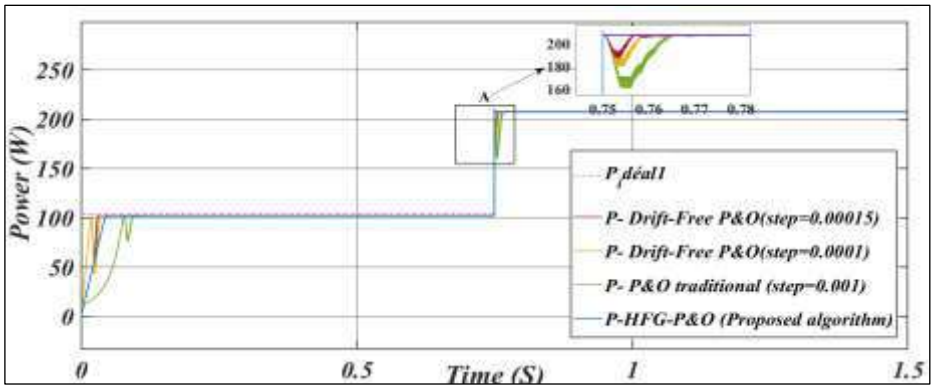


Figure 10: Extracted power comparison during Scenario 1.
Source: Authors, (2026).

Figure 11 reveals how step size affects the transient response: larger steps like 0.00015 yield faster convergence but cause noticeable oscillations near the MPP. HFG-P&O resolves this trade-off by using AGD to speed up tracking smoothly, avoiding overshoot. Figure 12 confirms that HFG-P&O also performs best in steady-state, limiting oscillations to only 0.16 W significantly lower than the 0.43 W and 0.44 W seen with conventional and Drift-Free P&O. This represents a 63% reduction in power ripple.

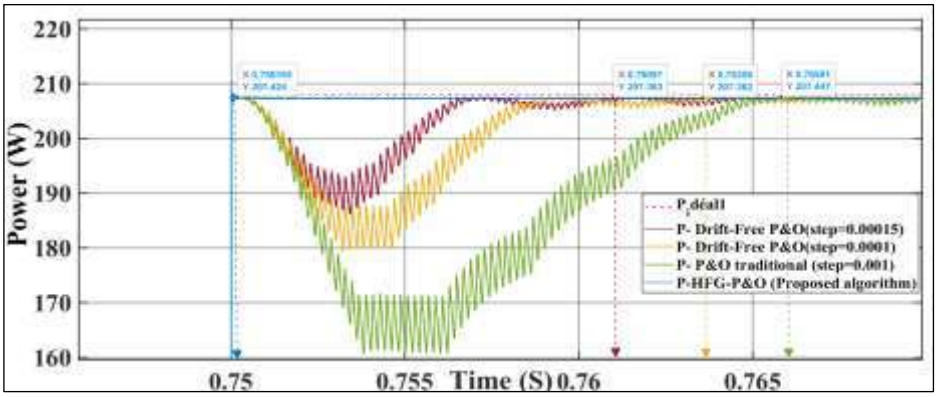


Figure 11: Transient response characteristics during irradiance transition (Point A).
Source: Authors, (2026).

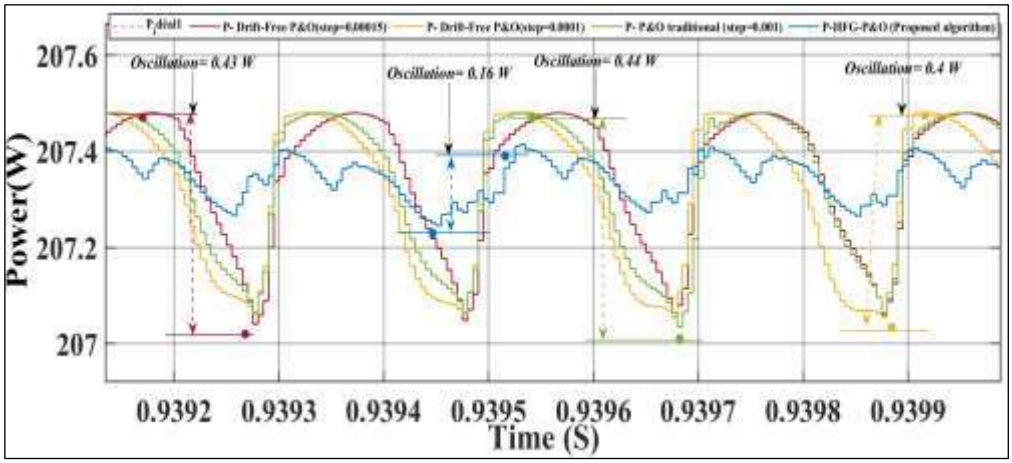


Figure 12: Steady-state power oscillation comparison at MPP.
Source: Authors, (2026).

Figure 13 shows the performance of four MPPT algorithms under highly dynamic irradiance (250–1000 W/m²). All experienced power drops, but the HFG-P&O consistently outperformed others by combining fuzzy logic and AGD for real-time step-size adjustment. It achieved fast convergence and stable power output. Drift-Free P&O with a small step was stable but slow, while the larger step improved speed at the expense of oscillations. Conventional P&O had the weakest response, with slow tracking and high losses due to its fixed step size.

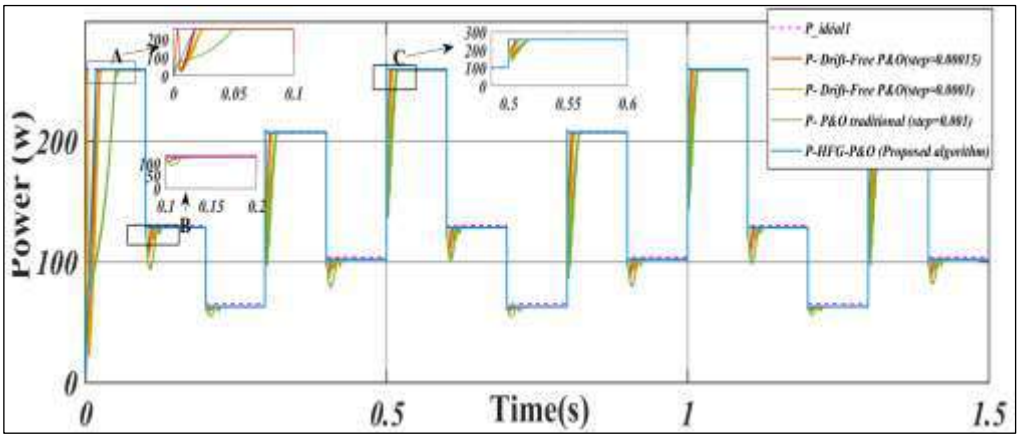


Figure 13: Algorithm performance under random irradiance fluctuations.
Source: Authors, (2026).

Figure 14 and Figure 15 illustrate how the algorithms react to a sudden irradiance rise from 250 W/m² to 1000 W/m² (Point A). HFG-P&O reached the MPP fastest (0.012 s) with low oscillations (0.3 W). Drift-Free P&O followed at 0.017 s (1.3 W ripple for larger step size), while the smaller step size version was more stable but slower (0.022 s). Traditional P&O was the slowest (0.05 s) and most unstable (>2.5 W ripple), confirming HFG-P&O's superior speed and stability.

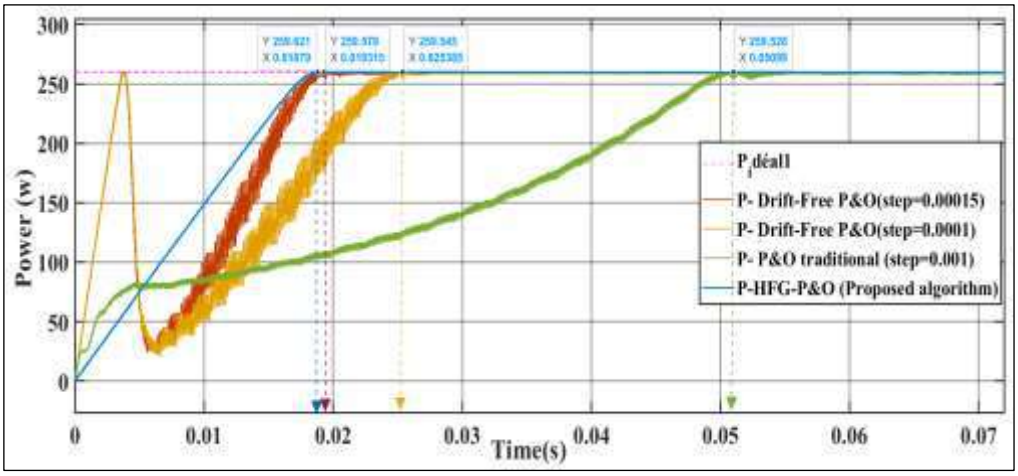


Figure 14: Dynamic response of MPPT algorithms at point A.
Source: Authors, (2026).

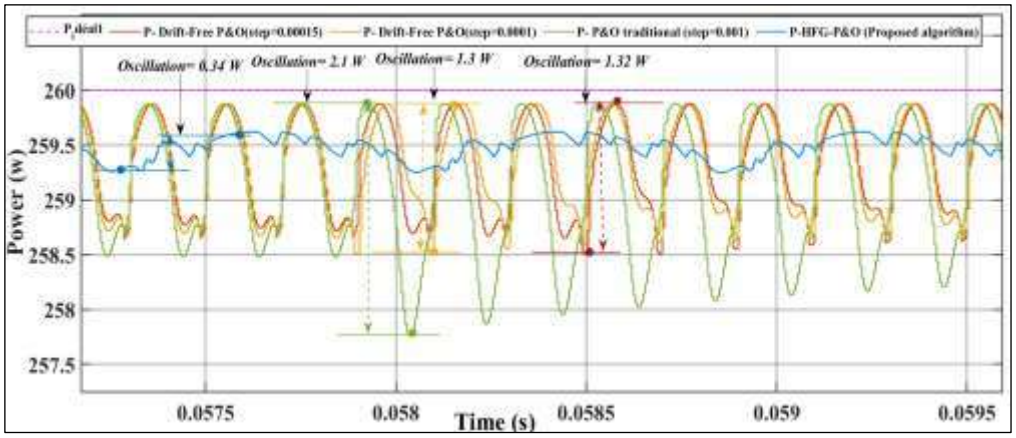


Figure 15: Dynamic response of MPPT algorithms at point A.
Source: Authors, (2026).

Figure 16 and Figure 17 illustrate the system's response to a sudden irradiance drop from 1000 W/m² to 500 W/m² (Point B). HFG-P&O reacted immediately at 0.1 s and stabilized in just 0.02 s with minimal oscillations (0.05 W). Drift-Free P&O with a 0.00015 step responded at 0.12 s but showed higher ripples (0.45 W), while the 0.0001 version was slower (0.025 s) with moderate oscillations (0.38 W). Traditional P&O was the slowest and most unstable, settling at 0.13 s with 0.42 W oscillations. These results confirm HFG-P&O's superior speed and stability under sudden irradiance drops.

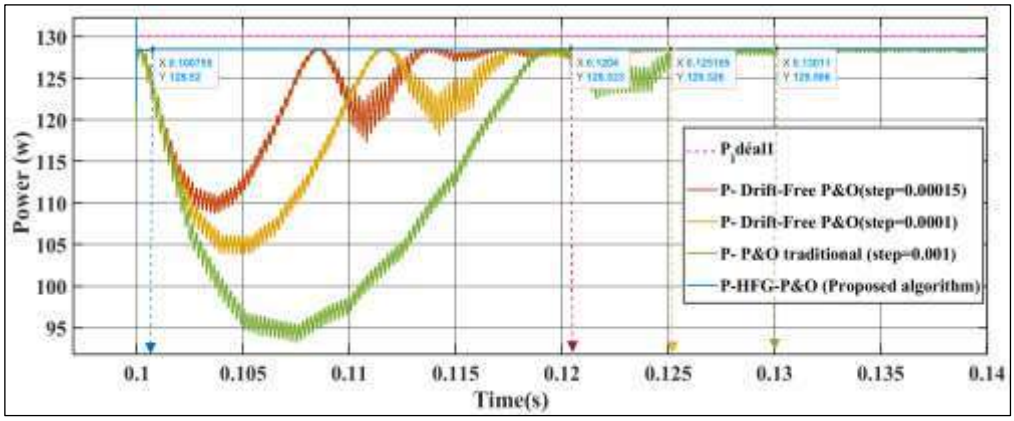


Figure 16: Dynamic response of MPPT Algorithms to a sudden drop in irradiance from 1000 W/m2 to 500 W/m2 at point B.
Source: Authors, (2026).

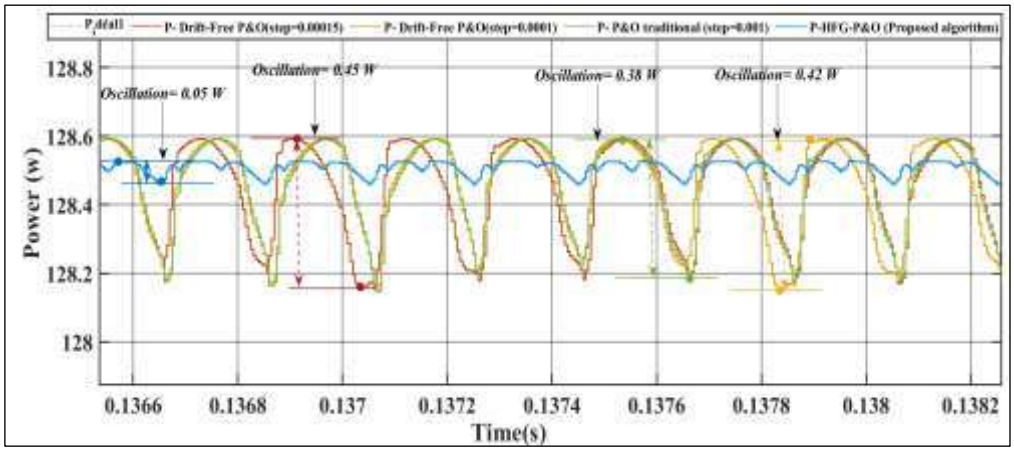


Figure 17: Post-stabilization oscillations in MPPT algorithms following an irradiance reduction at point B.
Source: Authors, (2026).

As shown in Figure 18 and Figure 19, during the second irradiance surge from 400 W/m² to 1000 W/m² (Point C), the HFG-P&O algorithm delivered the fastest and most stable response reacting at 0.5 s, stabilizing in 0.018 s, with only 0.34 W oscillations. In comparison, Drift-Free P&O (0.00015) settled in 0.022 s but exhibited 1.32 W ripples, while the 0.0001 version took 0.026 s with 1.3 W oscillations. Traditional P&O showed the slowest response (0.035 s) and the highest instability (> 2.1 W), confirming that HFG-P&O ensures faster and smoother MPPT even under steep irradiance changes.

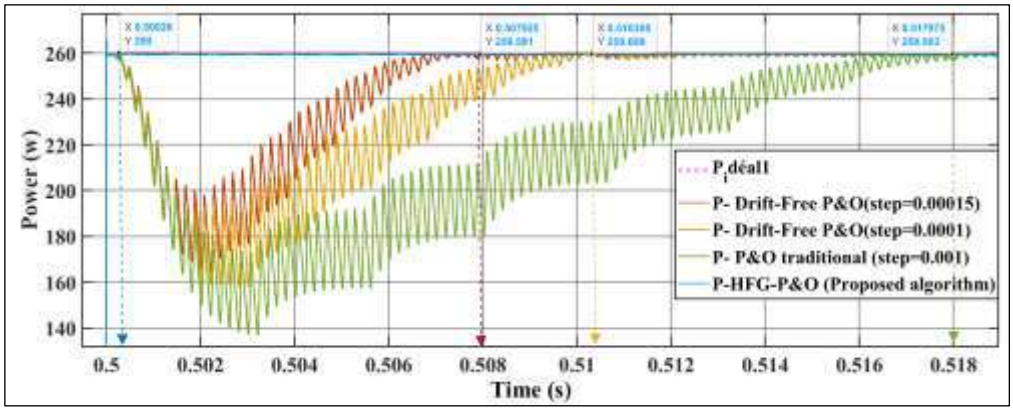


Figure 18: Power response of MPPT algorithms to a rapid increase in solar irradiance at point C.
Source: Authors, (2026).

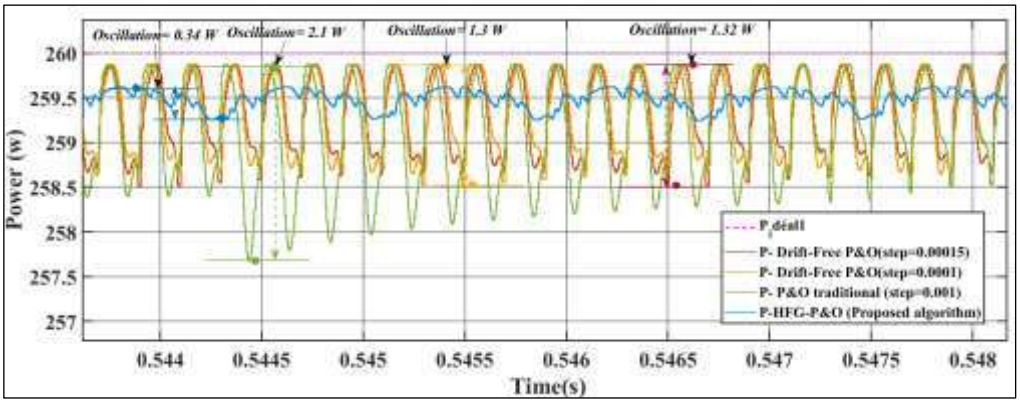


Figure 19: Post-stabilization power oscillations of MPPT algorithms at point.
Source: Authors, (2026).

Although based on simulations, the scenario closely mimics real-time PV dynamics, validating the practical potential of the proposed HFG-P&O MPPT algorithm. Table 4 compares MPPT methods under different irradiance shifts in terms of response time, stability, oscillations, overshoot, and efficiency. Using Equation 6, efficiency is calculated based on the deviation between actual power and the ideal MPP. HFG-P&O achieved the best result, with 98.5% efficiency, outperforming both Drift-Free and conventional P&O by maintaining precise and stable tracking under variable conditions.

Table 4: MPPT algorithm performance across different test scenarios.

Algorithm	**Scenario 1 – Point A**	**Scenario 2 – Point A**	**Scenario 2 – Point B**	**Scenario 2 – Point C**	**Overall Efficiency (%)**
HFG-P&O	Fastest response (0.0001s), quick stabilization, lowest oscillation (0.16W), minimal overshoot	Stable response (0.02s), minimal oscillations (0.3W), smooth adaptation	Immediate response (0.018s), lowest oscillation (~0.05W), minimal overshoot	Fastest response (0.012s), quick stabilization, lowest oscillation (0.34W), minimal overshoot	98.5%
Drift-Free P&O (0.00015)	Good response (0.0109s), higher oscillations (0.43W), moderate overshoot	Moderate response (0.025s), noticeable oscillations (1.3), higher overshoot	Acceptable response (0.022s), significant oscillations (0.45W), higher overshoot	Moderate response (0.025s), noticeable oscillations (1.32W), higher overshoot	96.67%
Drift-Free P&O (0.0001)	Slower response (0.0.0135s), moderate oscillations (~0.4W), higher overshoot	Delayed adaptation (0.03s), moderate oscillations (1.3W), overshoot	Slowest response (0.026s), lower oscillations (0.42W), prolonged stabilization	Slower response (0.022s), moderate oscillations (1.3W), higher overshoot	94.8%
Conventional P&O	Slowest response (0.0158s), high oscillations (0.4W), large overshoot, unstable	Significant delay (0.035s), high oscillations (2W), severe overshoot.	Slowest adaptation (0.035s), prolonged oscillations (0.38W), not suitable for dynamic conditions	Slowest response (0.05s), high oscillations (2.1W), large overshoot, unstable	93.2%

Source: Authors, (2026).

IV. CONCLUSIONS

This study set out to address the well-known limitations of conventional and modified MPPT techniques, namely, tracking drift, fixed step-size constraints, and poor adaptability under dynamic solar conditions as introduced in the opening section. In line with these objectives, the proposed Hybrid Fuzzy-Gradient Perturb and Observe (HFG-P&O) algorithm successfully demonstrated its ability to deliver real-time, drift-free, and oscillation-minimized tracking of the Maximum Power Point (MPP) across a variety of simulated

irradiance scenarios. Through extensive MATLAB/Simulink simulations, the HFG-P&O algorithm achieved its intended goals by dynamically adjusting the perturbation step size using a synergistic combination of fuzzy logic and Adaptive Gradient Descent (AGD). The results confirmed significant improvements in convergence speed, power stability, and tracking accuracy, achieving up to 98.5% efficiency, surpassing both classical P&O and Drift-Free P&O counterparts. These outcomes validate the initial expectations outlined in the Introduction and are thoroughly supported by the in-depth Results and Discussion analysis. The findings of this work establish HFG-P&O as a reliable, intelligent, and scalable MPPT solution, particularly advantageous for PV systems operating under rapidly changing environmental conditions. Future extensions of this research may include experimental validation, integration with predictive AI models for anticipatory control, and application under partial shading conditions to further validate the algorithm's generalizability and resilience.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Mahdi Bennoui, Ali Boukhari, Chouaib Labiod, Redha Meneceur, Abdelmoumene Hechifa, Ali Bebboukha.
Methodology: Mahdi Bennoui, Ali Boukhari, Chouaib Labiod, Redha Meneceur, Abdelmoumene Hechifa, Ali Bebboukha.
Investigation: Mahdi Bennoui, Ali Boukhari, Chouaib Labiod, Redha Meneceur, Abdelmoumene Hechifa, Ali Bebboukha.
Discussion of results: Mahdi Bennoui, Ali Boukhari, Chouaib Labiod, Redha Meneceur, Abdelmoumene Hechifa, Ali Bebboukha.
Writing – Original Draft: Mahdi Bennoui, Ali Boukhari, Chouaib Labiod, Redha Meneceur, Abdelmoumene Hechifa, Ali Bebboukha.
Writing – Review and Editing: Mahdi Bennoui, Ali Boukhari, Chouaib Labiod, Redha Meneceur, Abdelmoumene Hechifa, Ali Bebboukha.
Resources: Mahdi Bennoui, Ali Boukhari, Chouaib Labiod, Redha Meneceur, Abdelmoumene Hechifa, Ali Bebboukha.
Supervision: Mahdi Bennoui, Ali Boukhari, Chouaib Labiod, Redha Meneceur, Abdelmoumene Hechifa, Ali Bebboukha.
Approval of the final text: Mahdi Bennoui, Ali Boukhari, Chouaib Labiod, Redha Meneceur, Abdelmoumene Hechifa, Ali Bebboukha.

VI. REFERENCES

[1] Katche, M. L., Makokha, A. B., Zachary, S. O., & Adaramola, M. S. (2023). A Comprehensive Review of Maximum Power Point Tracking (MPPT) Techniques Used in Solar PV Systems. Energies, 16(5), 2206. https://doi.org/10.3390/en16052206.

[2] Singh, A. K., & Singh, R. R. (2021). An Overview of Factors Influencing Solar Power Efficiency and Strategies for Enhanci ng. 2021 Innovations in Power and Advanced Computing Technologies (i-PACT), 1–6. https://doi.org/10.1109/i-pact52855.2021.9696845.

[3] Mustafa, R. J., Gomaa, M. R., Al-Dhaifallah, M., & Rezk, H. (2020). Environmental Impacts on the Performance of Solar Photovoltaic Systems. Sustainability, 12(2), 608. https://doi.org/10.3390/su12020608.

[4] Dadkhah, J., & Niroomand, M. (2021). Optimization Methods of MPPT Parameters for PV Systems: Review, Classification, and Comparison. Journal of Modern Power Systems and Clean Energy, 9(2), 225–236. https://doi.org/10.35833/mpce.2019.000379.

[5] Hanzaei, S. H., Gorji, S. A., & Ektesabi, M. (2020). A Scheme-Based Review of MPPT Techniques With Respect to Input Variables Including Solar Irradiance and PV Arrays' Temperature. IEEE Access, 8, 182229–182239. https://doi.org/10.1109/access.2020.3028580.

[6] Ahmed, J., & Salam, Z. (2016). A Modified P&O Maximum Power Point Tracking Method With Reduced Steady-State Oscillation and Improved Tracking Efficiency. IEEE Transactions on Sustainable Energy, 7(4), 1506–1515.
https://doi.org/10.1109/tste.2016.2568043.

[7] Mahmod Mohammad, A. N., Mohd Radzi, M. A., Azis, N., Shafie, S., & Atiqi Mohd Zainuri, M. A. (2020). An Enhanced Adaptive Perturb and Observe Technique for Efficient Maximum Power Point Tracking Under Partial Shading Conditions. Applied Sciences, 10(11), 3912. https://doi.org/10.3390/app10113912.

[8] Singh, Shivendra, et al. "Implementation of Perturb & Observe MPPT Technique Using Bo ost Converter in PV System." 2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE), IEEE, 2020, pp. 1–4. DOI.org (Crossref), https://doi.org/10.1109/CISPSSE49931.2020.9212203.

[9] Sarvi, Mohammad, and Ahmad Azadian. "A Comprehensive Review and Classified Comparison of MPPT Algorithms in PV Systems." Energy Systems, vol. 13, no. 2, May 2022, pp. 281–320. DOI.org (Crossref), https://doi.org/10.1007/s12667-021-00427-x.

[10] Yadav, Vineet Kumar, et al. "Comparative Study of Different Variable Step Size Perturb and Observe Based MPPT." 2020 International Conference on Advances in Computing, Communication & Materials (ICACCM), IEEE, 2020, pp. 272–77.DOI.org(Crossref), https://ieeexplore.ieee.org/document/9212944.

[11] Killi, Muralidhar, and Susovon Samanta. "Modified Perturb and Observe MPPT Algorithm for Drift Avoidance in Photovoltaic Systems." IEEE Transactions on Industrial Electronics, vol. 62, no. 9, Sept. 2015, pp. 5549–59. DOI.org (Crossref), https://doi.org/10.1109/TIE.2015.2407854.

[12] Gopahanal Manjunath, Madhu, et al. "Adaptive Step Size Based Drift-free P&O Algorithm with Power Optimiser and Load Protection for Maximum Power Extraction from PV Panels in Stand-alone Applications." IET Renewable Power Generation, vol. 15, no. 6, Apr. 2021, pp. 1270–85. DOI.org (Crossref), https://doi.org/10.1049/rpg2.12105.

[13] Yung Yap, Kah, et al. "Artificial Intelligence Based MPPT Techniques for Solar Power System: A Review." Journal of Modern Power Systems and Clean Energy, vol. 8, no. 6, 2020, pp. 1043–59. DOI.org (Crossref), https://doi.org/10.35833/MPCE.2020.000159.

[14] Alhusseini, Haider, et al. "A Fuzzy-Based Adaptive P&O MPPT Algorithm for PV Systems With Fast Tracking and Low Oscillations Under Rapidly Irradiance Change Conditions." IEEE Access, vol. 12, 2024, pp. 84374–86. DOI: 10.1109/ACCESS.2024.3412848.

[15] Robles Algarín, Carlos, et al. "Fuzzy Logic Based MPPT Controller for a PV System." Energies, vol. 10, no. 12, Dec. 2017, p. 2036. DOI.org (Crossref), https://doi.org/10.3390/en10122036.

[16] S, Nivedha, and Vijayalaxmi. M. "Performance Analysis of Fuzzy Based Hybrid MPPT Algorithm for Photovoltaic System." 2021 International Conference on Communication, Control and Information Sciences (ICCISc) [Idukki, India], 2021, pp. 1–4. DOI: 10.1109/ICCISc52257.2021.9484882.

[17] Zerouali, Mohammed, et al. “Improvement of Conventional MPPT Techniques P&O and INC by Integration of Fuzzy Logic.” 2019 7th International Renewable and Sustainable Energy Conference (IRSEC), IEEE, 2019, pp. 1–6. DOI.org (Crossref), DOI: 10.1109/IRSEC48032.2019.9078330.

[18] Chou, Kuan-Yu, et al. “Adaptive Neuro Fuzzy Inference System Based MPPT Algorithm Applied to Photovoltaic Systems Under Partial Shading Conditions.” 2020 International Automatic Control Conference (CACS), IEEE, 2020, pp.16.DOI.org(Crossref), DOI: 10.1109/CACS50047.2020.9289733.

[19] Kuate Nkounhawa, Pascal, et al. “MPPT Based on Adaptive Neuro-Fuzzy Inference System (ANFIS) for a Photovoltaic System Under Unstable Environmental Conditions.” American Journal of Energy Engineering, vol. 9, no. 3, 2021, p. 68. DOI.org (Crossref), DOI:10.11648/j.ajee.20210903.12.

[20] Masry, Mohamed Zaghloul-El, et al. “New Hybrid MPPT Technique Including Artificial Intelligence and Traditional Techniques for Extracting the Global Maximum Power from Partially Shaded PV Systems.” Sustainability, vol. 15, no. 14,
July 2023, p. 10884. DOI.org (Crossref), https://doi.org/10.3390/su151410884.

[21] Kamoona, Mustafa A., et al. “Advanced MPPT Strategy for PV Solar Energy Systems Using ANN-GWO.” 2024 6th Global Power, Energy and Communication Conference (GPECOM), IEEE, 2024, pp. 439–44. DOI.org (Crossref), https://doi.org/10.1109/GPECOM61896.2024.10582686.

[22] Xu, Zhiwei, et al. “Photovoltaic Maximum Power Point Tracking Technology Based on Power Prediction Algorithm Combined with Variable Step Length Disturbance Observation Method.” Distributed Generation & Alternative Energy Journal, Oct. 2024, pp. 851–74. DOI.org (Crossref),https://doi.org/10.13052/dgaej2156-3306.3948.

[23] Raj, S. Antony, and G. Giftson Samuel. “Survey of AI Based MPPT Algorithms in PV Systems.” 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), IEEE, 2022, pp. 597–604. DOI.org (Crossref), https://doi.org/10.1109/ICSSIT53264.2022.9716521.

[24] Manna, Saibal, et al. “Hybrid Two-stage Adaptive Maximum Power Point Tracking for Stand-alone, Grid Integration, and Partial Shaded PV System.” International Journal of Adaptive Control and Signal Processing, vol. 37, no. 12, Dec. 2023, pp. 3297–327. DOI.org (Crossref), https://doi.org/10.1002/acs.3684.

[25] Manna, Saibal, et al. “Novel Lyapunov-Based Rapid and Ripple-Free MPPT Using a Robust Model Reference Adaptive Controller for Solar PV System.” Protection and Control of Modern Power Systems, vol. 8, no. 1, Apr. 2023, p. 13. BioMed Central, https://doi.org/10.1186/s41601-023-00288-9.

[26] Abbas, Ali, et al. “Power Tracking and Performance Analysis of Hybrid Perturb–Observe, Particle Swarm Optimization, and Fuzzy Logic-Based Improved MPPT Control for Standalone PV System.” Technologies, vol. 13, no. 3, Mar.2025,p.112. DOI: 10.3390/technologies13030112.

[27] Mhanni, Youssef, and Youssef Lagmich. “Adaptive Metaheuristic Strategies for Optimal Power Point Tracking in Photovoltaic Systems under Fluctuating Shading Conditions.” EPJ Photovoltaics, vol. 15, 2024, p. 31. www.epj-pv.org, https://doi.org/10.1051/epjpv/2024026.

[28] Ortiz-Munoz, Diana, et al. “Hybrid Fuzzy–DDPG Approach for Efficient MPPT in Partially Shaded Photovoltaic Panels.” Applied Sciences, vol. 15, no. 9, Apr. 2025, p. 4869. DOI.org (Crossref), https://doi.org/10.3390/app15094869.

[29] Pandey, A., et al. “High-Performance Algorithms for Drift Avoidance and Fast Tracking in Solar MPPT System.” IEEE Transactions on Energy Conversion, vol. 23, no. 2, Jun. 2008, pp. 681–89. DOI.org (Crossref), DOI: 10.1109/TEC.2007.914201.

[30] Zhang, Xiaofeng, et al. “Self-Optimized Drift-Avoidance MPPT Algorithm for Photovoltaic Systems.” 2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT) [Dali, China], 2022, pp.739–43. DOI.org(Crossref), https://ieeexplore.ieee.org/document/9986947.

[31] Zainuri, Muhammad Ammirrul Atiqi Mohd, et al. “Development of Adaptive Perturb and Observe-fuzzy Control Maximum Power Point Tracking for Photovoltaic Boost Dc–Dc Converter.” IET Renewable Power Generation, vol. 8, no. 2, Mar. 2014, pp. 183–94. DOI.org (Crossref), https://doi.org/10.1049/iet-rpg.2012.0362.

[32] Sera, D., et al. “Optimized Maximum Power Point Tracker for Fast-Changing Environmental Conditions.” IEEE Transactions on Industrial Electronics, vol. 55, no. 7, July 2008, pp. 2629–37. DOI.org (Crossref), https://doi.org/10.1109/TIE.2008.924036.

[33] Khabou, H., et al. “MPPT Implementation on Boost Converter by Using T–S Fuzzy Method.” Mathematics and Computers in Simulation, vol. 167, Jan. 2020, pp. 119–34. DOI.org (Crossref), https://doi.org/10.1016/j.matcom.2018.05.010.

[34] Rani, S., Giridhar, S., & Prasad, S. (2018). Effect of temperature and irradiance on solar module performance. IOSR Journal of Electrical and Electronics Engineering, 13(2), 36-40.

[35] Femia, N., et al. “Optimizing Duty-Cycle Perturbation of P&O MPPT Technique.” 2004 IEEE 35th Annual Power Electronics Specialists Conference (IEEE Cat. No.04CH37551), IEEE, 2004, pp. 1939–44. DOI.org (Crossref), https://ieeexplore.ieee.org/document/1355414.

[36] Rico-Camacho, Roberto I., et al. “Transient Differentiation Maximum Power Point Tracker (Td-MPPT) for Optimized Tracking under Very Fast-Changing Irradiance: A Theoretical Approach for Mobile PV Applications.” Applied Sciences, vol. 12, no. 5, Mar. 2022, p. 2671. DOI.org (Crossref), https://doi.org/10.3390/app12052671.

[37] Waghmare, V. B., and C. Nikate Swapnali. “A Drift Free Perturb & Observe MPPT in PV System.” 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), IEEE, 2017, pp. 1–5. DOI.org (Crossref), https://doi.org/10.1109/ICECCT.2017.8118022.

[38] Pilakkat, Deepthi, and S. Kanthalakshmi. “Drift Free Variable Step Size Perturb and Observe MPPT Algorithm for Photovoltaic Systems Under Rapidly Increasing Insolation.” Electronics ETF, vol. 22, no. 1, Aug. 2018, p. 19. DOI.org (Crossref), https://doi.org/10.7251/ELS1822019P.

[39] Manna, Saibal, and Ashok Kumar Akella. “Comparative Analysis of Various P & O MPPT Algorithm for PV System under Varying Radiation Condition.” 2021 1st International Conference on Power Electronics and Energy (ICPEE), IEEE, 2021, pp. 1–6. DOI.org (Crossref), https://doi.org/10.1109/ICPEE50452.2021.9358690.

[40] D. Sera, T. Kerekes, R. Teodorescu and F. Blaabjerg, "Improved MPPT Algorithms for Rapidly Changing Environmental Conditions," 2006 12th International Power Electronics and Motion Control Conference, Portoroz, Slovenia, 2006, pp. 1614-1619 https://doi.org/10.1109/EPEPEMC.2006.4778635.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 661-670. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2957

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

TO3: A COMPACT AND EFFICIENT ENCODING SCHEME FOR SKEWED SLICING FLOORPLANS IN VLSI DESIGN

Biswojit Nayak*[1] and Mrutynjaya Panda[2]

[1, 2]Department of Computer Science and Applications, Utkal University, Vani Vihar, Bhubaneswar- 758044, Odisha, India.

[1]https://orcid.org/0000-0002-7912-6065, [2]https://orcid.org/0000-0001-5713-9220

Email: *biswojit22@gmail.com, mrutyunjaya.cs@utkaluniversity.ac.in

ARTICLE INFO

Article History
Received: November 16, 2025
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
VLSI,
Slicing Floorplan,
Slicing Tree,
TO3 Encoding.

ABSTRACT

A slicing floorplan is a recursive process of drawing a floorplan horizontally and vertically until all the blocks are accommodate into it. In the nanometer era, the escalating transistor density in VLSI chips underscores the necessity of designing compact floorplan representations for VLSI circuits. This work introduces a novel approach, employing a tree of order three (TO3) for the effective coding of a skewed slicing tree that corresponds to a slicing floorplan. The key innovation lies in the utilization of a TO3 structure, which adds a distinctive dimension to the encoding scheme. The proposed encoding method for a skewed slicing tree achieves a notable reduction in code size compared to previous approaches. Experimental findings from the MCNC benchmark and artificial circuits indicate that the proposed code uses, on average, almost 80% fewer bits than Breadth First code (BFS), 46% fewer than Slicing Pair (SP) code, and 49% fewer than Improved Slicing Pair (ISP) code for encoding a slicing tree, while also reducing CPU running time through a more compact representation of the electronics modules.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Floorplan design is one of the most significant phases in VLSI physical design, which involves arranging a collection of rectangular circuit modules on a chip to optimize the overall area and the length of interconnection wires. This arrangement defines the layout of a VLSI circuit, which is a well-known NP-hard problem [1] that has attracted a lot of interest in recent years. Floorplans are generally divided into two categories: slicing floorplans and non-slicing floorplans. A slicing floorplan is created by recursively drawing horizontal and vertical lines from a given rectangle as shown in Figure 1 (a). In contrast, a non-slicing floorplan cannot be recursively divided into two halves to obtain all of the blocks of the given rectangle. Non-slicing floorplans are sometimes known as spiral or wheel floorplans, as shown in Figure 1 (b).

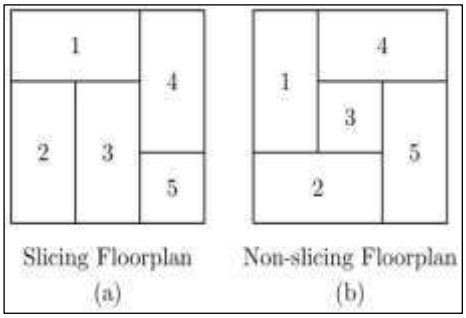


Figure 1: Slicing and non-slicing floorplans.
Source: Authors, (2026).

A slicing floorplan is commonly represented using a binary tree known as a slicing tree. In this representation, each internal node is labelled with '+' or '*' based on line drawing types (for horizontal lines, '+' and for vertical lines, '*'), and leaf nodes represent electronics blocks.

However, slicing trees are not unique representations of the floorplan since the same layout can often be represented by different slicing trees depending on how the lines are chosen. To address this issue, [2] introduced the concept of a skewed slicing tree in 1986, in which the parent and the right child are always opposite line drawing types, as illustrated in Fig 2. However, the non-slicing floorplan is a generalized floorplan such as the one shown in Fig. 1 (b), can be represented in several alternative models, including Sequence Pair (SP), O-Tree, Transitive Closure Graph (TCG), Corner Block List (CBL), SKB-tree, and B*-tree [3-7].

Later, in 2001, [8] proved that a slicing tree can serve as a complete representation for general floorplans. They also showed that, when combined with a simple compaction process, slicing tree representations are capable of producing maximally compact module placements. The proposed work focuses on slicing floorplans because of their hierarchical structure and suitability for machine learning-based generation, as demonstrated in recent work by [9]. Building on this idea, the slicing tree is presented using a binary encoding scheme, enabling faster computation and reduced storage compared to symbolic tree traversal used in [9].

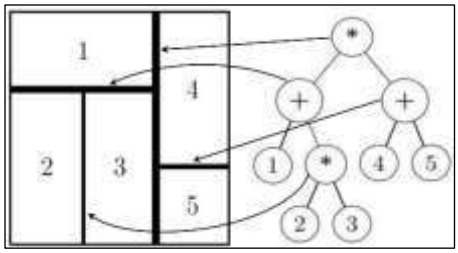


Figure 2: Slicing floorplan and its corresponding skewed slicing tree.
Source: Authors, (2026).

Over the years, several binary representations of floorplans have been proposed in the literature [10-17]. In 2006, [10] proposed a $(5n-5)$ bits representation of a floorplan, where n denotes the number of inner rectangles. Later [11] introduced a more compact $(4n-4)$-bits representation of floorplan and also showed that the number of different rectangles satisfies $R(n) <= 13.5(n-1)$. In 2012, [12] proposed two compact encodings of size $6f-2n^4+6$ bits and $5f-5$ bits for floorplans, where n represents the number of edges, f is the number of inner faces, and n^4 is the number of vertices of degree four. The optimum encoding is dependent on the number of vertices of degree four. The first encoding is more efficient when $2n^4 > f+11$, while the second performs better otherwise. In a subsequent work, [13] proposed an even more compact encoding of size $5f-n^4-6$ bits for $f >= 2$ of general rectangular drawings, which is independent of n^4.

To represent the mosaic (dead space free) floorplan, [14] introduced a string-based data structure known as the quarter-state sequence or Q sequence. The number of bits needed to represent mosaic floorplans with n blocks using the Quarter-State Sequence approach is 4n bits. Subsequently, He [15] proposed more compact binary encoding of length $(3n-3)$ bits and also proved that any binary string representation of an n block mosaic floorplan must use at least $(3n-o(n))$ bits. In 2014, [16] introduced two encoding schemes for slicing floorplans: the breadth-first code (BFS code) and the slicing-pair code (SP code). For a slicing floorplan with n blocks, the BFS code requires $3n-2$ bits. Building on this idea, the authors introduced the SP code by improving the BFS code.

The length of the SP code is $5n/2+m_1-p_{10}-4$ bits, Where n denotes the number of blocks, m_1 represents the number of internal nodes in a slicing tree that have exactly one child as a leaf node (i.e., a block), and p_{10} counts the number of slicing pairs (siblings) in which the left node is a leaf and the right node is an internal node. More recently, [17] proposed an improved slicing-pair (ISP) code designed for a balanced skewed slicing floorplan. They also proved that the ISP code has a length of $2(n+m_1)-3$ bits. Meanwhile, [9] proposed a subitizing-inspired approach to VLSI floorplanning that fine-tunes large language models (LLMs) to generate slicing trees representing near-optimal floorplans. Their method encodes the layout as post-order traversals of slicing trees and achieves high success and low dead-space ratios, especially when using GPT-4o-mini.

Unlike their post-order symbolic representation, the present work introduces a binary encoding of slicing floorplans, providing a more compact and machine-interpretable form that facilitates efficient optimization and learning. The motivation for creating compact representations of VLSI floorplans stems from the necessity of being able to quickly optimize very large and complex designs using minimal computational resources. Compact representations also require less memory to store and optimize the floorplan, which is particularly advantageous for large scale designs. Thus, even complex designs can be optimized with minimal resources as well as reduced execution times due to quicker evaluation of the design space during optimization processes [18]. The perturbation technique allows for an efficient examination of the design space without having to recreate the entire floorplan, thus speeding up the optimization process [14], [18].

In addition, compact encodings lead to faster runtime of both the floorplan generation and verification process, which are especially important in large industrial designs with a tight schedule for market entry [19]. Although hybrid placement methods such as [20] have shown promising results, they still require significant computational effort for their search processes. The reinforcement learning-based optimization technique proposed by [21] reduces the burden of explorations by learning efficient perturbations of sequence-pair floorplans but still depends on the compactness and flexibility of the underlying binary encoding. The use of a more compact encoding could help lower the complexity of that search by reducing the number of bits required to represent the solution space. Building on this motivation, the current work improves existing solutions such as SP code [16] and ISP code [17].

The present contribution introduces a new approach to encode skewed slicing floorplans referred to as the Tree of Order Three (TO3) encoding. To summarize, this paper makes the following key contributions.

- A tree of order three (TO3) code is introduced to encode skewed slicing floorplans.

- On average, the length of TO3 code reduces the length by 80%, 46%, and 49% compared to BFS, SP, and ISP codes, respectively.
- The range of the TO3 code length (i.e., lower and upper bounds) is determined based on [16] and [17], and its performance is compared with the lengths of the SP and ISP codes.
- Interestingly, the upper bound of the TO3 code length matches the lower bound of the SP and ISP codes, thus proving the TO3 code is more compact than the SP and ISP codes.

The remainder of this paper is organized as follows. Section II reviews the existing BFS, SP, and ISP encoding schemes. Section III introduces the proposed TO3 code for skewed slicing floorplans. The implementation details are discussed in Section IV, and Section V concludes with a summary of findings and direction for future work.

II. PRELIMINARIES

This section presents essential definitions and important finding from state-of-the-art research, which collectively guide the development of the TO3 code discussed in Section III.

II.1 DEFINITIONS

II.1.1 Slicing Floorplan

The floorplan consisting of exactly one block is referred to as a slicing floorplan. A slicing floorplan F with n blocks can be constructed by drawing horizontal and vertical cuts recursively on each slicing floorplan that partitions floorplan into two smaller slicing floorplans, F_1 and F_2. This partition process continues until all the blocks fit into the floorplan. In case of a vertical cut, the right-most boundary of F_1 is equivalent to the left-most boundary of F_2, while a horizontal cut aligns the top-most horizontal boundary of F_1 is equivalent to the bottom-most boundary of F_2.

II.1.2 Slicing Tree

A slicing floorplan F consisting of n blocks is represented by a binary tree known as a slicing tree T with $2n-1$ vertices or nodes. In this tree, every internal node represents either '*' (for vertical cut) or '+' (for horizontal cut), and the leaf node correspond to the electronics blocks. A slicing tree is called skewed if an internal node and its right child are of opposite cuts, as shown in Fig. 2; otherwise, the tree is non-skewed.

II.1.3 Slicing Pair

A slicing pair (V_i, V_{i+1}) for each $i = 1,3,\ldots,2n-3$, consists of two vertices in the slicing tree T (with n blocks) that have the same parent P, where $(V_i, V_{i+1}) \in \{*, +, L\}$, $P \in \{*, +\}$, and L represents a leaf node, as illustrated in Figure 3.

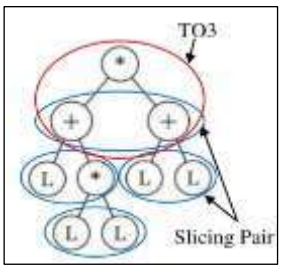


Figure 3: Slicing pairs and TO3 representation of a skewed slicing tree T.
Source: Authors, (2026).

II.2 PRIOR WORKS

Theorem 1. For a slicing floorplan F with n blocks, the BFS representation, i.e., $|B(F)|$, requires $3n-2$ bits. In addition the floorplan can be encoded and decoded in linear time, that is, $O(n)$ [16].

Theorem 2. Consider a slicing floorplan F with n blocks and its associated skewed slicing tree T. If |SP(F)| represents the number of bits required to encode F using the slicing pair code, then the encoding length is $|SP(F)| = 5n/2 + m_1/2 - p_{10} - 4$ bits, Where m_1 and p_{10} are same parameters introduced in Section I. Moreover, both encoding and decoding using slicing pair code runs in $O(n)$ time [16].

Theorem 3. The length of $SP(F)$ is bounded by the inequality $2n-3 \leq |SP(F)| \leq 3n-5$. When the constant terms in these expressions are ignored, the worst case of SP and BFS codes are almost the same [16].

Theorem 4. Consider a slicing floorplan F with n blocks and its corresponding skewed slicing tree is T. Let $|ISP(F)|$ represent the length of the encoded bits string of T using the improved slicing pair code, i.e., $|ISP(F)| = 2(n + m_1) - 3$, where m_1 is the number of internal vertices that have exactly one leaf node [17].

Theorem 5. For a floorplan F with n blocks, if the corresponding skewed slicing tree is balanced, then $|ISP(F)| < |SP(F)| < |B(F)|$ holds[17].

III. PROPOSED TO3 CODE FOR SKEWED SLICING FLOORPLAN

This section presents a novel encoding technique specially designed for skewed slicing floorplans. It explains the underlying principle of encoding. In addition, this section presents detailed derivation of the code length and the range length of the proposed code.

III.1 ENCODING

Consider a slicing floorplan F with n blocks and let T(F) be its associated skewed slicing tree. A tree of order three, in short TO3 (i.e., a subtree with three vertices) is obtained form T(F), which is represented as a triplet (x, y, z). In this triplet, $x \in \{+,*\}$ represents the root node, while $y, z \in \{+,*, L\}$, denotes the left and right children of x respectively. The rules to obtain TO3s of a skewed slicing tree are outlined below.

1 Internal vertices at even levels of the slicing tree, starting from the root (considering the root at level 0) are candidates for roots of TO3s.
2 If the root of a TO3 is at level i, then two leaves of the TO3 are at level $(i + 1)$, and they can't be roots of any other TO3s, i.e., TO3s are disjoint trees.
3 A leaf labelled by L can not be the root of a TO3.

Using the above rules and the property of a skewed slicing tree (an internal node always has a label opposite to that of its right child), the possible TO3s for $x = '+'$ of the type $(+, y, z)$ are: $(+, +,*)$, $(+, +, L)$, $(+,*,*)$, $(+,*, L)$, $(+, L,*)$, $(+, L, L)$. Similarly, TO3s of the type $(*, y, z)$ are: $(*, +, +)$, $(*, +, L)$, $(*,*, +)$, $(*,*, L)$, $(*, L, +)$, $(*, L, L)$. In the skewed slicing tree T shown in Figure 3, the corresponding TO3 is inscribed within a circle. The TO3 code, analogous to the SP and ISP codes, is shown in Table 1. Column 1 of Table 1 represents the TO3 labels, and the encoded bit pattern used for each TO3 label is shown in Column 2. Although L is not a TO3 label, it is included in the last row of the table and assigned a bit value of 1, as the label L is required to encode when it is followed by TO3 labels in the TO3's level-wise traversal. The complete encoding procedure for the TO3 code is described in Algorithm 1.

Table 1: Patterns of TO3 and their assigned codes.

TO3 Labels	Assigned Bits
(+, +, *)	011
(+, +, L)	010
(+, *, *)	001
(+, *, L)	000
(+, L, *)	01
(+, L, L)	00
(*, +, +)	111
(*, +, L)	110
(*, *, +)	101
(*, *, L)	100
(*, L, +)	11
(*, L, L)	10
L	1

Source: Authors, (2026).

Algorithm 1 takes the root of the skewed slicing tree as its input and generates the corresponding encoded bit string. The algorithm employs a queue data structure Q to traverse the skewed slicing tree level-wise, during which all the TO3s are identified and encoded using the bit pattern listed in Table 1. Since Algorithm 1 visited every node of the slicing tree exactly once, and a slicing tree with n leaves contains $n - 1$ internal nodes, the overall time complexity of the algorithm is $O(n)$. By applying bit patterns from Table 1 and following Algorithm 1, the TO3 code length for the skewed slicing tree shown in Fig. 3 is found to be 6 bits. In comparison, the lengths of the BFS [16], SP [16], and ISP [17] codes for the same tree are 13 bits, 8 bits, and 9 bits, respectively. Therefore, among the BFS, SP, ISP, and TO3 codes, the proposed TO3 code required the fewest number of bits and provide the most compact.

Algorithm 1: Encoding.

Input: Root of the skewed slicing tree
Output: Encoded String (EncodeStr)
1: Initialize EncodeStr ← Null
2: Insert root of the skewed slicing tree into a queue Q (Enqueue(Q, root))
3: while Q ≠ Ø do
4: if Front → ch == L then
5: Append encoded TO3 label (Front → ch) using Table 1 to EncodeStr
6: else
7: Append encoded TO3 label (Front → ch, Front → Left → ch, Front → Right → ch) using Table 1 to EncodeStr
8: end if
9: if Front → Left → Left ≠ NULL then
10: Enqueue(Q, Front → Left → Left)
11: end if

```
12:   if Front → Left → Right ≠ NULL then
13:      Enqueue(Q, Front → Left → Right)
14:   end if
15:   if Front → Right → Left ≠ NULL then
16:      Enqueue(Q, Front → Right → Left)
17:   end if
18:   if Front → Right → Right ≠ NULL then
19:      Enqueue(Q, Front → Right → Right)
20:   end if
21:   Dequeue(Q)
22: end while
23: return EncodeStr
```

Source: Authors, (2026).

III.2 DECODING

The decoding algorithm described in Algorithm 2 takes an encoded binary string as input and reconstructs the corresponding skewed slicing tree. The bit(s) assigned (in Table 1) for the TO3 label either start with 0 (for the parent of the TO3 label is '+') or start with 1 (for the parent of the TO3 label is '*'). An encoded bit string starting with '0' has an assigned length of either three or two bits, whereas a bit string starting with '1' may have a length of three, two, or one bit. Algorithm 2 is a recursive algorithm that makes decisions non-deterministically based on the current symbol pointed to by the forward pointer (FP). During execution the algorithm generates TO3 labels level-wise and stores them in a queue data structure Q. Since Algorithm 2 non-deterministically generates TO3 labels, there is the possibility of more than one slicing tree, which is held by an array named ArraySkewed. Algorithm 2 then systematically builds and validates all slicing trees using skewed slicing properties, ultimately selecting a valid skewed slicing tree as the output. In the worst case, the slicing tree has n leaf nodes; the running time complexity of Algorithm 2 is $O(3^{(2n-1)})$, since the algorithm explore non-deterministically on bits directed by the forward pointer (FP).

Algorithm 2: Decoding.

```
Input: Encoded String (EncodeStr)
Output: Skewed Slicing Tree
1: Initialize forward pointer FP to the first character of the input encoded string
2: Initialize a queue Q to hold level-wise TO3 labels
3: Create an array named ArraySkewed to store the skewed slicing tree
4: Call AlgoDecode(EncodeStr, FP, Q, ArraySkewed)
5: if FP ≠ Null then
6:    if FP = 1 then
7:       TempStr = FP
8:       Decode TempStr to the TO3 label using Table 1
9:       Append TO3 label to Q
10:   AlgoDecode(EncodeStr, FP + 1, Q, ArraySkewed)
11:   end if
12:   if (FP = 1 ∨ FP = 0) ∧ (FP + 1 ≠ Null) then
13:      Append two bits from FP to TempStr
14:      Decode TempStr to the TO3 label using Table 1
15:      Append TO3 label to Q
16:      AlgoDecode(EncodeStr, FP + 2, Q, ArraySkewed)
17:   end if
18:   if (FP = 1 ∨ FP = 0) ∧ (FP + 2 ≠ Null) then
19:      Append three bits from FP to TempStr
20:      Decode TempStr to the TO3 label using Table 1
21:      Append TO3 label to Q
22:      AlgoDecode(EncodeStr, FP + 3, Q, ArraySkewed)
23:   end if
24: else
25:   Append Q to ArraySkewed
26:   return ArraySkewed
27: end if
28: Generate level-wise slicing trees on the array of queues and check whether
the tree is skewed or not
29: return Skewed Slicing Tree
```

Source: Authors, (2026).

III.3 ESTIMATING LENGTH OF TO3 CODE

This subsection derives an expression for the length of the TO3 code. The analysis follows the same approach used to determine the length of the SP code in [16]. To support this derivation, the following additional notations are considered for the skewed slicing tree T(F), as illustrated in Table 2.

Table 2: Notations and Definitions for TO3 Encoding.

Notation	Definition
|TO3(F)|	Length of TO3 code of T(F) in bits for a floorplan F.
t_{000}	Number of TO3s (x, y, z) such that x, y, z ∈ {+, *}.
t_{001}	Number of TO3s (x, y, z) such that x, y ∈ {+, *}, z ∈ {L}.
t_{010}	Number of TO3s (x, y, z) such that x, z ∈ {+, *}, y ∈ {L}.
t_{011}	Number of TO3s (x, y, z) such that x ∈ {+, *}, y, z ∈ {L}.
m	Number of internal nodes of T(F).
p	Number of leaf nodes (labeled by L) of T(F) that are not part of any TO3.

Source: Authors, (2026).

For the skewed slicing tree shown in Fig. 3, $p = 3$, there are three such Ls at level 2 of T. The expression for the bit length of the TO3 code corresponding to the slicing tree T(F) is given in Theorem 6.

Theorem 6. Let T(F) be a skewed slicing tree for slicing floorplan F with n blocks. Then the length of the TO3 code for encoding T(F) is $2n - t_{010} - t_{011} - 3$ bits. i.e., $|TO3(F)| = 2n - t_{010} - t_{011} - 3$.

Proof. Since the slicing tree T(F) is a complete binary tree with n blocks, it contains $m = n - 1$ internal nodes. To derive an expression for m, observe that each TO3 of the type t_{000} has three internal nodes, TO3s of the types t_{001} and t_{010} have two internal nodes, and each TO3 of the type t_{011} has one internal node and two leaf nodes. Additionally, the expression for the number of leaf nodes p that are not part of the TO3 labels is given in Eq. 2. Thus, it follows that

$$m = n - 1 = 3t_{000} + 2t_{001} + 2t_{010} + t_{011} \quad (1)$$

$$p = n - t_{001} - t_{010} - 2t_{011} \quad (2)$$

To establish an expression for |TO3(F)|, referencing Table 1 reveals that TO3s of types t_{000} and t_{001} require 3 bits each, while types t_{010} and t_{011} utilize 2 bits each. Additionally, the leaf L, which is not part of any TO3, is represented with just one bit. In summary, this information contributes to the derivation of the total bit count for |TO3(F)| as present in Eq. 3.

$$\begin{aligned} |TO3(F)| &= 3t_{000} + 3t_{001} + 2t_{010} + 2t_{011} \\ &+(n - t_{001} - t_{010} - 2t_{011} - 2) \end{aligned} \quad (3)$$

Recall that in the calculation of |SP(F)| (See Theorem 2, Section II) the authors of [16] saved the last two bits for slicing pair (L, L) subtracting 2 from the expression of |SP(F)|. Similarly, in order to have fair comparison between |SP(F)| and |TO3(F)|, last two bits for slicing pair (L, L) is also saved by subtracting 2 from the right hand side of Eq. 3. By substituting Eq. 2 into Eq. 3, the resulting expression is derived.

$$\begin{aligned} |TO3(F)| &= n + 3t_{000} + 2t_{001} + t_{010} - 2 \\ &= n + 3t_{000} + 2t_{001} + 2t_{010} - t_{010} \\ &\qquad + t_{011} - t_{011} - 2 \\ &= n + (3t_{000} + 2t_{001} + 2t_{010} + t_{011}) \\ &\qquad - t_{010} - t_{011} - 2 \\ &= n + n - 1 - t_{010} - t_{011} - 2 \\ &= 2n - t_{010} - t_{011} - 3 \end{aligned} \quad (4)$$

Equation 4 estimates the bit-length of the TO3 code of T(F). Hence the proof. The lengths of the BFS, SP, ISP, and TO3 codes are computed for the skewed slicing tree in Fig. 3. For the parameter values $n = 5, m_1 = 1, p_{10} = 1, t_{010} = 0, t_{011} = 1$, Theorem 1, 2, 4 and 7 yields $|B(F)| = 3 * 5 - 2 - 4 = 13$ bits, $|SP(F)| = 12.5 + 0.5 - 1 - 4 = 8$ bits, $|ISP(F)| = 2 * (5 + 1) - 3 = 9$ bits and $|TO3(F)| = 10 - 0 - 1 - 3 = 6$ bits respectively. Hence, the proposed TO3 code encodes T(F) using a smaller number of bits compared to the other codes.

III.4 RANGE LENGTH OF TO3 CODE

The authors of [16] showed that, in the worst case, the length of SP code is almost the same as that of the BFS code, indicating that SP offers a more compact representation (see Theorem 3, Section II). Further, [17] proved that, for balanced skewed slicing trees, the ISP code achieves better compression than the SP code (see Theorem 5, Section II). Extending these results, the following section demonstrates that the worst-case length of the TO3 code is shorter than the best-case length of the ISP and SP codes. In summary, the following Theorem 7 is established.

Theorem 7. If T(F) be the skewed slicing tree of a floorplan F with n blocks. Then $\frac{3n-5}{2} \leq |TO3(F)| \leq 2n - 3$.

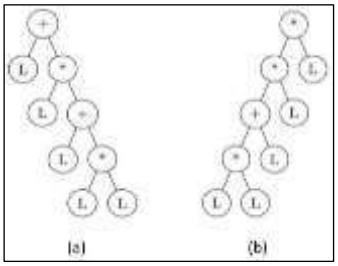


Figure 4: Minimum and Maximum saved skewed slicing trees.
Source: Authors, (2026).

Proof. The best and worst case length of TO3 code are evaluated using the skewed slicing trees shown in Fig. 4 (a) and Fig. 4 (b). The best case of $|TO3(F)|$ occurs when the slicing tree skews along the right branch of the tree shown in Fig. 4 (a). In this case: for n: odd, $t_{010} = \frac{n-1}{2}$, $t_{011} = 0$, and for n: even, $t_{010} = \frac{n-2}{2}$, $t_{011} = 1$. Using these values in $|TO3(F)| = 2n - t_{010} - t_{011} - 3$, the following result is obtained.

$$|TO3(F)| = \frac{3n-5}{2} \quad \text{if } n : \text{odd}$$
$$= \frac{3n-6}{2} \quad \text{if } n : \text{even}$$

Therefore, the tight lower bound of $|TO3(F)|$ is $\frac{3n-5}{2}$. i.e.

$$\frac{3n-5}{2} \leq |TO3(F)| \tag{5}$$

Similarly, for the worst case of $|TO3(F)$ occurs when the slicing tree skews along the right branch of the tree shown in Fig. 4 (b). In such case: for n: odd, $t_{010} = 0$, $t_{011} = 0$, and for n: even, $t_{010} = 0$, $t_{011} = 1$. Using the values of t_{010} and t_{011} in $|TO3(F)| = 2n - t_{010} - t_{011} - 3$, it follows that

$$|TO3(F)| = 2n - 3 \quad if\ n : odd$$
$$= 2n - 4 \quad if\ n : even$$

Thus the tight upper bound of $TO3(F)$ is $2n - 3$. i.e
]

$$|TO3(F)| \leq 2n - 3 \tag{6}$$

From inequality 5 and inequality 6, it follows $\frac{3n-5}{2} \leq |TO3(F)| \leq 2n - 3$. □
Following Theorem 3, 5, and 7, Corollary 1 can be derived.

Corollary 1: $\frac{3n-5}{2} \leq |TO3(F)| \leq 2n - 3 \leq |SP(F)|$ or $|ISP(F)| \leq 3n - 5$
Thus, To3 code is more compact than BFS, SP and ISP codes.

III.5 EXPECTED NUMBER OF SPs AND TO3s

The length of encoded bits required for a slicing floorplan depends on the number of slicing pairs (SPs) or TO3s present in the corresponding slicing tree. This section presents results on the expected number of SPs and TO3s of a slicing tree with n blocks.

Theorem 8. For skewed slicing tree with n blocks and height $h = \log n$, the expected number of SPs is given by $\frac{4(4^h-1)-3h}{9(n-1)} \approx \frac{4n^2-3\log n-4}{9(n-1)}$.

Proof. Let X_i be a random variable that denotes the number of SPs at level $1 \leq i \leq h$ of the slicing tree. Let X be the random variable that denote the total number of SPs in the tree. Then $X = X_1 + X_2 + \ldots + X_h$. Clearly $X_i \in \{1,2,4,\ldots,2^{i-1}\}$, since the root of the tree does not participate in any slicing pair. The tree contains n leaves and $n-1$ internal nodes; hence the cardinality of the sample space is $n-1$. Let $E[X]$ and $E[X_i]$ are expectation of random variables X and X_i respectively. Then $E[X_i] = \frac{1}{n-1}\sum_{i=1}^{l}(2^{i-1})^2$. Thus, by linearity of expectation, and taking $h = \log n$,

$$E[X] = \sum_{l=1}^{h-1} E[X_i] = \frac{1}{3(n-1)} \sum_{l=1}^{h-1} 4^l - 1$$
$$= \frac{4(4^h-1)-3h}{9(n-1)}$$
$$\approx \frac{4n^2-3\log n-4}{9(n-1)}.$$

Theorem 9. Consider a skewed slicing tree with n blocks and height $h = \log n$, the expected number of TO3s is given by $\frac{16(4^{h+1}-1)-3h}{15(4^{\lceil \log_4 n \rceil}-1)}$. Here h is taken as a $h-1$ when h is odd and as $h-2$ when h is even.

Proof. Since nodes at even levels $0,2,4,\ldots m$ (where $m = h-1$ if h is odd and $m = h-2$ if h is even) of the slicing tree are roots of TO3s. Let $X_i (i \in \{0,2,4,\ldots m\})$ be a random variable that depresents the number of TO3s at level i and X be the random variable that denotes the total number of TO3sof the slicing tree, then, $X = X_0 + X_2 + X_4 \ldots + X_m$. The cardinality of the sample space is $\frac{4^{\lceil \log_4 n \rceil}-1}{3}$. Let $E[X]$ and $E[X_i]$, denote the expectations of random variables X and X_i respectively. Then, $E[X_i] = \frac{3}{4^{\lceil \log_4 n \rceil}-1} \sum_{i=1}^{\frac{m}{2}} (4^i)^2$. By linearity of expectation,

$$E[X] = \sum_{m=0}^{h} E[X_i] = \frac{1}{5(4^{\lceil \log_4 n \rceil}-1)} \sum_{m=0}^{h} 16^{\frac{m}{2}+1} - 1$$
$$= \frac{16(4^{h+1}-1)-3h}{15(4^{\lceil \log_4 n \rceil}-1)}$$

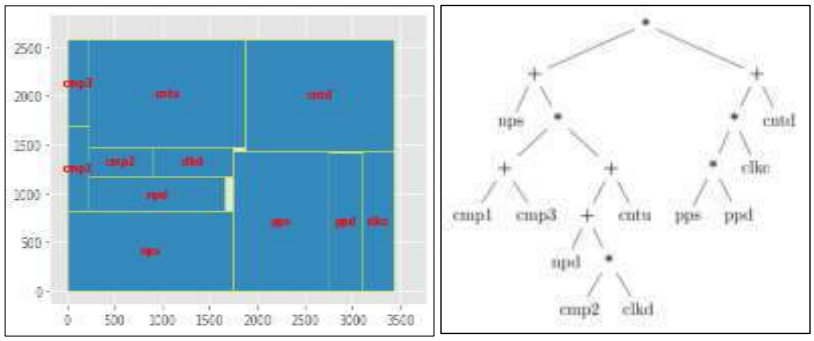
When the height of the slicing tree is odd, $h = h-1$ with $h = \log n$; the expected value is given by $E[X] \approx \frac{16n^2-3\log n-13}{15(n-1)}$. For slicing tree even height, $h = h-2$ with $h = \log n$; in this case, the expected value becomes $E[X] \approx \frac{4n^2-3\log n-10}{15(n-1)}$. □

In the worst case, where both a slicing pair (SP) and a TO3 are encoded using 3 bits. Theorem 8 and Theorem 9 shows that the number of bits required to encode the expected number of TO3s is less than that of SPs.

IV. IMPLEMENTATION AND RESULT ANALYSIS

In this section, the results of the proposed TO3 code are validated with state-of-the-art works using MCNC benchmarks [3] and artificial modules. The proposed and the other algorithms are implemented in Python and executed on a Windows 10 PC equipped with an AMD 3.2 GHz CPU and 8 GB RAM. The length of binary code of slicing trees corresponding to MCNC benchmark floorplans are validated using the B*-tree-based Simulated Annealing (BTBSA) algorithm. Since MCNC benchmark circuits are non-slicing in nature, each floorplan is transformed into a slicing floorplan using the compaction technique discussed in [8]. The resulting slicing trees are then encoded using BFS, SP, ISP, and TO3 coding techniques, which are tabulated in Table 3, which lists the benchmark names, number of modules, and the encoded bit lengths produced by each algorithm.

Figure 5 (a) shows the floorplan of the HP benchmark circuit generated by the BTBSA algorithm, and the corresponding slicing tree is illustrated in Figure 5 (b). For artificial modules, slicing trees are randomly generated for different numbers of modules, and the length of binary code of four encoding techniques is summarized in Table 3. The last row of Table 3 presents the comparison results, where the average values indicate the improvement achieved by the proposed TO3 coding scheme over BFS, SP, and ISP. For a fair comparison, the lengths of the state-of-the-art codes have been adjusted using the TO3 code. From the last row of Table 3, it can be observed that, on average, BFS, SP, and ISP codes need 80%, 46%, and 49% more bits, respectively, than TO3 codes for encoding a slicing tree.



(a) A floorplan generated on HP benchmark. (b) Slicing tree corresponding to floorplan Fig. 5 (a).
Figure 5: A floorplan and its corresponding slicing tree.
Source: Authors, (2026).

Additionally, Figure. 6 (a) and Figure 6 (b) present a comparison of the BFS, SP, ISP, and TO3 code lengths for benchmark and artificial circuits, respectively. The horizontal axis defines different benchmark or artificial circuits, and the vertical axis defines the length of BFS, SP, ISP, and TO3 codes. The vertical bars in red, gray, green, and blue correspond in length to the BFS, SP, ISP, and TO3 codes, respectively. The vertical bar in blue, representing the length of the TO3 code, has the least height among the other encoding schemes for different circuits, as shown in Figure 6. Therefore, compared to BFS, SP, and ISP codes, TO3 code is more compact and uses fewer bits to encode a slicing tree.

Table 3: Comparison of lengths of BFS, SP, ISP, and TO3 codes for benchmark and artificial circuits.

Benchmark / Artificial	No. of Modules / Leaves	BFS Code Length (bits)	SP Code Length (bits)	ISP Code Length (bits)	TO3 Code Length (bits)
ami33	33	97	83	97	57
ami49	49	145	125	153	86
hp	11	31	24	29	18
apte	9	25	19	25	14
xerox	10	28	19	25	14
AM1	5	13	8	9	4
AM2	10	28	21	17	17
AM3	20	58	46	37	27
AM4	35	103	83	69	58
AM5	49	145	118	97	78
Average		1.80	1.46	1.49	1.00

Source: Authors, (2026).

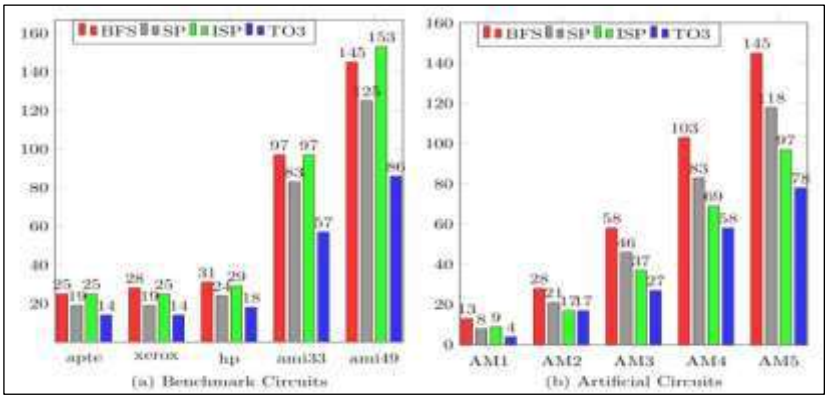


Figure 6: Lengths of BFS, SP, and TO3 codes for benchmark and artificial circuits.
Source: Authors, (2026).

The convergence iteration determines how effectively the algorithm explores and settles into an optimal or near-optimal floorplan layout, which shows an efficient search process and lower computational costs. As shown in Table 4, during the convergence iterations, the slicing tree-based simulated annealing (STBSA) with code representation yields experimental results comparable to those achieved without the code representation of BTBSA. In contrast, the CPU runtime is reduced significantly when the code representation is employed. Specifically, the runtime of STBSA on each benchmark, i.e., ami33, ami49, hp, apte, and xerox, is faster than that of BTBSA by 16%, 11%, 14%, 17%, and 8%, respectively. This improvement arises because the proposed strategy required fewer bits to represent the electronic blocks or modules, thereby reducing computational overhead without compromising the solution quality.

Table 4: Comparison results with and without the slicing tree-based simulated annealing.

Benchmark	Convergence Iterations (With)	Convergence Iterations (Without)	Time (With) (seconds)	Time (Without) (seconds)	Ratio
ami33	59	62	24.23	28.74	0.843
ami49	56	54	41.56	46.65	0.891
hp	19	20	15.52	18.03	0.861
apte	16	14	2.85	3.43	0.831
xerox	11	12	3.95	4.27	0.925

Source: Authors, (2026).

The TO3 code offers the most compact representation when compared with BFS, SP, and ISP, as demonstrated by both the theoretical bounds in Eq. 4 and the implementation results in Table 3. This compactness is consistently observed across small and large circuits, leading to less memory use and faster processing. A key advantage of the TO3 code is that it is faster in terms of CPU runtime because of its simpler structure and efficient encoding process. Since each TO3 is a disjoint subtree, modification or perturbation to any particular TO3 triplet affects only that triplet and its immediate neighbors, unlike in SP or BFS, where dependencies can spread across the entire structure. This disjoint property makes TO3 more efficient for both encoding and optimization tasks.

V. CONCLUSION

This paper proposed a compact binary encoding technique, namely TO3, for the skewed slicing tree that corresponds to a slicing floorplan, and derived an analytical expression for both the length and range length of the TO3 code. The theoretical analysis demonstrates that the expected number of TO3s is smaller than the number of slicing pairs (SPs) required to encode a slicing tree. More importantly, the experimental results on benchmark circuits and artificial modules show that the proposed coding scheme is 80%, 46%, and 49% more compact than the BFS, SP, and ISP codes, respectively. In addition, the code-based simulated annealing reduces CPU running time compared to the B*-tree-based simulated annealing. In the future, this work will be extended to non-slicing floorplans.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Biswojit Nayak and Mrutynjaya Panda.
Methodology: Biswojit Nayak.
Investigation: Biswojit Nayak and Mrutynjaya Panda.
Discussion of results: Biswojit Nayak and Mrutynjaya Panda.
Writing – Original Draft: Biswojit Nayak.
Writing – Review and Editing: Biswojit Nayak and Mrutynjaya Panda.
Resources: Biswojit Nayak.
Supervision: Mrutynjaya Panda.
Approval of the final text: Biswojit Nayak and Mrutynjaya Panda.

VII. REFERENCES

[1] B. Yang, Q. Xu, H. Geng, S. Chen, B. Yu, and Y. Kang, "Floorplanning with edge-aware graph attention network and hindsight experience replay," ACM Trans. Des. Autom. Electron. Syst., vol. 29, no. 5, pp. 1–17, 2024.

[2] D. Wong and C. Liu, "A new algorithm for floorplan design," in 23rd ACM/IEEE design automation conference, IEEE, 1986, pp. 101–107.

[3] Y.-C. Chang, Y.-W. Chang, G.-M. Wu, and S.-W. Wu, "B*-trees: A new representation for non-slicing floorplans," in Proceedings of the 37th annual design automation conference, 2000, pp. 458–463.

[4] P.-N. Guo, C.-K. Cheng, and T. Yoshimura, "An o-tree representation of non-slicing floorplan and its applications," in Proceedings of the 36th annual ACM/IEEE design automation conference, 1999, pp. 268–273.

[5] P.-N. Guo, T. Takahashi, C.-K. Cheng, and T. Yoshimura, "Floorplanning using a tree representation," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 20, no. 2, pp. 281–289, 2001.

[6] J.-M. Lin and Y.-W. Chang, "TCG: A transitive closure graph-based representation for non-slicing floorplans," in Proceedings of the 38th annual design automation conference, 2001, pp. 764–769.

[7] X. Hong et al., "Corner block list: An effective and efficient topological representation of non-slicing floorplan," in IEEE/ACM international conference on computer aided design. ICCAD-2000. IEEE/ACM digest of technical papers (cat. No. 00CH37140), IEEE, 2000, pp. 8–12.

[8] M. Lai and D. Wong, "Slicing tree is a complete floorplan representation," in Proceedings design, automation and test in europe. Conference and exhibition 2001, IEEE, 2001, pp. 228–232.

[9] C.-C. Yeh, S.-C. Lu, H.-L. Cho, Y.-C. Lin, and R.-B. Lin, "Subitizing-inspired large language models for floorplanning," in Proceedings of the symposium on assessment of synthesis and integration of mixed information (SASIMI 2025), Yuan Ze University, 2025, pp. 223–228.

[10] K. Yamanaka and S. Nakano, "Coding floorplans with fewer bits," IEICE transactions on fundamentals of electronics, communications and computer sciences, vol. 89, no. 5, pp. 1181–1185, 2006.

[11] T. Takahashi, R. Fujimaki, and Y. Inoue, "A (4 n- 4)-bit representation of a rectangular drawing or floorplan," in Computing and combinatorics: 15th annual international conference, COCOON 2009 niagara falls, NY, USA, july 13-15, 2009 proceedings 15, Springer, 2009, pp. 47–55.

[12] M. Saito and S. Nakano, "Two compact codes for rectangular drawings with degree four vertices," in Proc. 11th forum on information technology (FIT 2012), 2012, pp. 1–8.

[13] T. Takahashi, "A compact code for rectangular drawings with degree four vertices," Journal of information processing, vol. 22, no. 4, pp. 634–637, 2014.

[14] K. Sakanushi, Y. Kajitani, and D. P. Mehta, "The quarter-state-sequence floorplan representation," IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, vol. 50, no. 3, pp. 376–386, 2003.

[15] B. He, "Optimal binary representation of mosaic floorplans and baxter permutations," in Frontiers in algorithmics and algorithmic aspects in information and management: Joint international conference, FAW-AAIM 2012, beijing, china, may 14-16, 2012. proceedings, Springer, 2012, pp. 1–12.

[16] T. Ohmori, K. Yamanaka, T. Hirayama, and Y. Nishitani, "Compact codes of slicing floorplans," 電子情報通信学会技術研究報告= IEICE technical report: 信学技報, vol. 114, no. 80, pp. 33–37, 2014.

[17] B. Nayak and B. Ray, "ISP: An improved slicing pair code for skewed slicing floorplan," in 2023 36th international conference on VLSI design and 2023 22nd international conference on embedded systems (VLSID), IEEE, 2023, pp. 205–210.

[18] S. Vinay Kumar, P. Rao, and M. K. Singh, "Optimal floor planning in VLSI using improved adaptive particle swarm optimization," Evolutionary Intelligence, vol. 15, no. 2, pp. 925–938, 2022.

[19] B. Srinivasan and R. Venkatesan, "Multi-objective optimization for energy and heat-aware VLSI floorplanning using enhanced firefly optimization," Soft Computing, vol. 25, no. 5, pp. 4159–4174, 2021.

[20] V. V. Ignatyev, A. V. Kovalev, O. B. Spiridonov, V. M. Kureychik, A. S. Ignatyeva, and I. B. Safronenkova, "Hybrid genetic–paired–permutation algorithm for improved VLSI placement," ETRI Journal, vol. 43, no. 2, pp. 260–271, 2021, doi: 10.4218/etrij.2019-0412.

[21] S. Yu, S. Du, and C. Yang, "A deep reinforcement learning floorplanning algorithm based on sequence pairs," Applied Sciences, vol. 14, no. 7, p. 2905, 2024.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 671-683. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.2967

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

PREDICTING STUDENTS' CONCENTRATION IN COGNITIVE ACTIVITIES USING EEG AND DEEP LEARNING TECHNIQUES

Yogesh Patil[1], Snehal Mali[2], Harsha Gaikwad*[3], Sanil Gandhi[4], Manjushree Laddha[5] and Arvind Kiwelekar[6]

[1, 2, 3, 4, 5, 6]Department of Computer Engineering, Dr. Babasaheb Ambedkar Technological University, Lonere – 402 103, Raigad (M.S.), India.

[1]https://orcid.org/0009-0003-6128-5740, [2]https://orcid.org/0000-0003-2760-0009, [3]https://orcid.org/0000-0003-2760-0009
[4]https://orcid.org/0000-0003-0407-8975, [5]http://orcid.org/0000-0002-6265-1728, [6]https://orcid.org/0000-0002-3407-0221

Email: *harsha.gaikwad@dbatu.ac.in, ynpatil@dbatu.ac.in, smali@dbatu.ac.in, ssgandhi@dbatu.ac.in, mdladdha@dbatu.ac.in, awk@dbatu.ac.in

ARTICLE INFO

Article History
Received: November 18, 2025
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Electroencephalogram (EEG), Concentration Level Prediction, Deep Learning Algorithms, Cognitive Engagement

ABSTRACT

In an era of social media and online learning platforms, there are several opportunities for learning different technologies and topics that students do not easily understand. However, it also presents challenges by diverting students' attention, such as notifications, multitasking activities, advertisements, etc. Assessing students' level of focus during cognitive tasks is crucial and complex. This study evaluates students' cognitive engagement through various activities, including arithmetic calculations, reading technical articles, listening to technical podcasts, reading transcripts, browsing the internet, and engaging in relaxation exercises, utilizing EEG signals. Concentration levels are classified using deep learning algorithms, specifically Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Artificial Neural Networks (ANN). The performance of these algorithms is also evaluated based on metrics such as accuracy, F1 Score, precision, and loss.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

With the advent of technology, the utilization of mobile phones has witnessed a surge across all age groups. Research has revealed that social networking applications offer users shorter material and can significantly increase their dopamine levels [1]. The floods of dopamine induced by these spikes lead individuals to develop an addiction to rapidly consuming stuff, so negatively impacting their capacity to sustain and concentrate on specific jobs for prolonged durations and acquire new knowledge. It is vital to evaluate the cognitive involvement of an individual when performing various jobs. This will help us determine their degrees of attentiveness to different cognitive focus requirements.

Within the realm of education, the ability to discern pupils' degrees of concentration during the acquisition of new knowledge might potentially enable us to intervene to sustain their attention [2]. The concentration level of students can be inspected by examining their eye contact, body language, and response to a question or communication feedback [3]. The concentration examination techniques are subjective and can be influenced by the student's mood or the external environment. Additionally, different students may have different concentration indicators, so applying one metric to all students will be accountable for inconsistency. It isn't easy to manually assess the concentration level of a student while performing any cognitive task.

To address these issues, a data-driven, objective process is needed to assess concentration levels. EEG is a promising solution for quantifying cognitive focus and engagement [4]. EEG captures electrical signals caused by neuronal communication in the brain, representing the fine-granular, real-time cognitive states. Initially, the EEG device is used for medical diagnosis, for example, for diagnosing epilepsy [5], Parkinson's disease [6], Alzheimer's disease [7], sleep disorders [8], and emotion detection [9], [10]. EEG is now being seriously examined in non-medical fields, such as in education [11–13], human-computer interfaces [10], and cognitive workload estimation [14], [15].

The success of previous studies in classifying cognitive states [14], human emotions [9], and comprehension levels [11], [16] in learning exercises using EEG signals is a proven track record. Cognitive engagement is directly linked to brain activity; EEG can provide a more timely and accurate reading of levels of focus than a test of behaviour [17]. When humans perform different cognitive

efforts, corresponding EEG patterns have different properties [16]. The complexity of EEG signals makes them challenging to analyze, as they are highly non-linear, multi-dimensional, and susceptible to noise. Although traditional machine learning algorithms have been proven capable of doing well in cognitive state classification based on EEG, it is hampered by their hand-engineered feature dependency [18]. The feature engineering process in analyzing EEG is traditionally complex and domain-specific, making scaling difficult.

The traditional algorithms also struggle to capture deep hierarchical representations in EEG signals [19], leading to a non-optimal outcome in concentration prediction. Deep learning is a revolutionizing tool for processing EEG signals, overtaking traditional machine learning algorithms [19]. The combined usage of ANNs and LSTMs in a joint model is highly accurate in cognitive state decoding, making it highly suitable for predictive levels of concentrations [16]. The intersection of deep learning and the processing of EEG signals is opening new paths for real-time monitoring of concentrations in learning, workplaces, and clinics [20]. In online learning contexts, EEG-based prediction of concentration can support adaptive learning systems reacting in a timely manner to learners' cognitive states [11–13], [21].

By identifying periods of lower levels of attention, one can adopt a smart tutoring system for teaching and improve the motivation and retention of the students. In occupational contexts, EEG-based concentration monitoring can enhance productivity by making tailored suggestions for inter-burst periods, task reconfigurations, and workloads [22], [23]. Additionally, in clinical contexts, prediction of levels of concentration can assist in diagnosing, in an early stage, attention-related diseases such as Attention-Deficit/Hyperactivity Disorder (ADHD), allowing for directed treatments for enhanced cognitive well-being [24], [25]. While it is potentially valuable, several challenges exist for EEG-based concentration prediction. The signals in an EEG can vary significantly between different individuals, necessitating robust feature extraction algorithms and adaptation in a specific domain [26], [27]. Computative efficiency is critical for real-time usage, especially for a wearable EEG device [28].

Transfer learning, data augmentation, and attention mechanisms have been tried to enhance a model's adaptability and generality in different populations [12], [21]. This study examines the EEG data collected during five different tasks. Each of these activities necessitates different degrees of cognitive involvement from an individual, necessitating differing amounts of concentration [12]. We employ deep learning algorithms to categorize these signals according to their concentration categories. The rest of the article is organized in the following sections: The related work in detailed and background about the several terminologies used in the experiment is described in section 2, followed by a detailed description of the methodology in section 3. The performance measures used for the described methods are in section 4. The results of the proposed methodologies are discussed in section 5, and finally, the article concludes in section 6.

II. RELATED WORK AND BACKGROUND

This section reviews previous research work using EEG signals to predict concentration, attention level, and cognitive load using deep learning algorithms. The authors [19] presented a comprehensive review of recent advancements in deep learning methodologies for classifying EEG signals. The paper also elucidated the efficacy of various deep learning architectures, including ANN, CNN, LSTM, and MLP, in the processing and analysis of EEG data and compares deep learning approaches with the traditional machine learning techniques, highlighting the superior accuracy and efficiency achieved through deep learning in EEG classification tasks. The authors [16] preferred EEG recordings to classify students' cognitive states during learning activities automatically. The data preprocessing techniques included noise reduction and artefact removal from raw EEG signals, followed by feature extraction using summary statistics from various frequency bands.

Several classification algorithms were used in the study, with logistic regression serving as the primary technique for model training. These algorithms include local models, global models, and multi-task learning approaches. According to the data, global models—which take into account more contextual information—predict recall levels more accurately than local models, which predict concentration levels with an accuracy of 91.88%. To address the impact of distractions on cognitive performance, the authors used EEG signals to classify human concentration levels [21], [29]. The dataset comprises brain signals collected using an EEG device from individuals in both relaxation and concentration states. This data was preprocessed using high-pass filtering to eliminate low-frequency noise and segmentation of signals into the delta, theta, alpha, and beta frequency bands.

Statistical features were calculated using the PyEEG library to extract features. The classification of concentration levels is executed using the Support Vector Machine (SVM) algorithm with a radial basis function (RBF) kernel, achieving an overall accuracy of 84% to classify EEG signals into low and high concentration levels. The research findings of [30] suggested that effective brain-machine interfaces can be developed for distinguishing mental states, even with inexpensive EEG devices. The authors used a low-cost, toy-grade EEG device to collect EEG signals to classify relaxed and concentrated mental states. Data was preprocessed by converting raw EEG signals into time-frequency representations and aggregating spectral components into EEG bands such as alpha, beta, and gamma.

The SVM and multilayer feedforward neural networks (BPNN) classification algorithms were preferred in the study. Experimental results demonstrated that using features from the combined alpha, beta, and gamma bands with a bandwidth of 4Hz yields an average accuracy exceeding 80% across subjects, with some individuals achieving accuracies above 90% using the SVM classifier. The authors put forward a strong correlation between students' concentration levels and learning outcomes by analysing brain signals captured by an EEG device [31], [32]. The brain wave data was collected from junior high school students performing cognitive tasks, specifically the Culture Fair Intelligence Test (CFIT) and the Indonesian Competency Test (CT). EEG signals were extracted using Fast Fourier Transform (FFT) followed by normalization to standardize the data range.

The study used the K-Nearest Neighbor (KNN) algorithm, with a K-value of 5 for identifying patterns in the extracted data and reports an impressive classification accuracy of 94.59% to classify concentration levels. The classification accuracy was evaluated using K-Fold Cross Validation with a K-value of 11. The study by [22] investigated young adults' (ages 18-30) cognitive activity and concentration levels using EEG data. At the same time, participants performed IQ tests based on the Standard Progressive Matrices

(SPM). The data preprocessing involved signal reduction and filtering beta waves in the frequency range of 13-30Hz to enhance the data quality. The classification of concentration levels was performed using the k-nearest Neighbor (KNN) algorithm into three categories: high, medium, and low. The model achieved an accuracy of 70%, providing insights into the correlation between brain activity and cognitive performance. The paper presented by [33] gave a method for classifying EEG signals into three mental states: relaxation, moderate, and high concentration.

The brain signals were captured using an EEG device with four electrodes and preprocessed by applying a large Laplacian filter, a notch passive filter, and a wavelet filter. Mental states in the time domain were examined by segmenting the measured signal into windows, from which various features are extracted. These extracted signals were utilized in machine learning algorithms to classify the different mental states. The two-fold cross-validation method was used to evaluate the classification. A novel model tailored for embedded systems to classify attention levels while performing specific tasks. This system used a single-channel EEG device to capture brain signals that were subsequently increased through spectrum analysis to increase the dimensionality of the features. The authors used simulated annealing and geometric particle swarm optimization algorithms to optimize feature selection. Classification of mental states (concentrated or relaxed) was achieved using an SVM. Furthermore, Echo State Networks, a class of neural networks well-suited for embedded systems, demonstrated satisfactory accuracy in meeting the model's requirements [34].

The research paper [13] tackled the critical issue of assessing student attention in online learning, a challenge intensified by the COVID-19 epidemic. The authors proposed a brain-computer interface (BCI) system that uses EEG data to evaluate student involvement and attention levels. A publicly available EEG dataset was preferred, and the quick Fourier transform extracted power spectral density (PSD) features and various attention metrics were calculated. The study encompassed three classification algorithms: KNN, SVM, and random forest (RF), with the RF model achieving the highest accuracy of 96% in holdout validation. The method encompassed data collection, preprocessing, feature extraction, and model training, culminating in classification performance assessment. The results indicated that the suggested RF model successfully differentiates attention states, presenting a viable alternative for improving online learning experiences.

The classification of normal and pathological EEG patterns using deep learning models enhanced with attention mechanisms is investigated in the paper [26]. A systematic approach was used to preprocess EEG data to filter and normalize signals, extract time-domain and frequency-domain features, and segment the data into frames for study. The authors tested InstaGATs, LSTM with Attention, and CNNs with Attention against baseline models without attention mechanisms. The CNNs with Attention model has the highest accuracy of 96.98% on the TUH EEG Seizure dataset. The research faced intrinsic noise in EEG recordings, the difficulty of distinguishing anomalies, and the need for efficient feature extraction from varied datasets. Attention processes can improve classification performance by stressing relevant data items and enhancing clinical EEG analysis interpretability and efficacy. Attention-enhanced deep learning models can significantly improve EEG classification tasks, enabling more robust implementations in real applications.

II.1 BACKGROUND

II.1.1 EEG Device

The neuron is the smallest unit of the human brain. Electric signals are generated when neurons communicate with each other. An electroencephalogram is a device used to capture the electrical signals generated by neurons in the brain. These electrical signals are recorded with the electrodes of an EEG device. With the help of electric signals captured by EEG devices, abnormal changes in the brain can be detected, which is why EEG devices are widely used in the medical sector to diagnose brain-related diseases such as epilepsy, sleep disorders, and so on. In recent years, EEG devices have also been utilized in the education sector to solve various research problems, such as determining attention levels, mental engagement, and cognitive load. This study uses a MUSE EEG device to record brain wave data while performing various cognitive tasks. This device has four electrodes named FP1 and FP2, located in the frontal lobe of the brain, and TP9 and TP10, located in the parietal lobe of the brain, as shown in Figure 1.

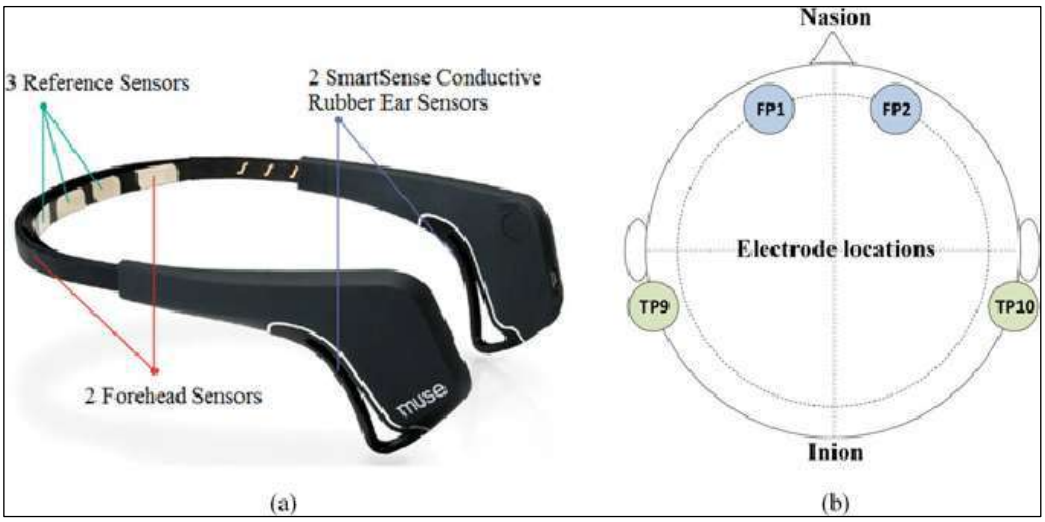


Figure 1: MUSE EEG Device.
Source: Authors, (2026).

Table 1: Types of Brain Waves.

Types of Wave	Frequency (Hz)	Task
Delta	0.5 – 4	Deep sleep
Theta	4 – 8	Relaxing or meditating
Alpha	8 – 12	Just after woke up and before going to sleep
Beta	12 – 35	Routine work
Gamma	> 35	Deep concentration on some task or doing cognitive activity

Source: Authors, (2026).

II.1.2 EEG Signals

Five different types of brain waves are emitted by the brain and recorded using an EEG device, as shown in Table 1. Delta and theta waves are small amplitude waves emitted, implying an individual's deep sleep or relaxation. Their frequency ranges from 0.5Hz to 8Hz. These waves are produced in large quantities as compared to other types of waves while doing tasks such as relaxing with closed or open eyes, browsing the internet, and so on. When humans are in partial sleep, waking up, or transitioning into a deep sleep state, the brain emits alpha waves with a frequency range of 8Hz to 12Hz. The beta and gamma waves indicate high cognitive activity. Their frequency range is 12Hz to 35Hz and >35Hz, respectively. Among all the waves, gamma and beta waves are produced in large quantities with more intensity while doing high-concentration tasks such as doing arithmetic calculations or reading technical articles.

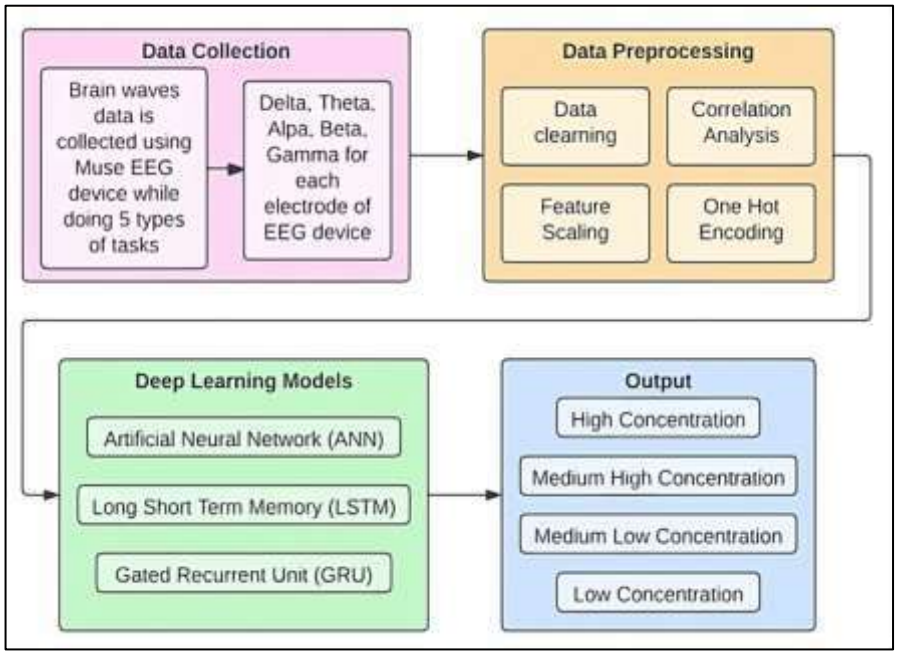


Figure 2: Flowchart Diagram of the Proposed System.
Source: Authors, (2026).

III. METHODOLOGY

The proposed methodology for predicting concentration levels utilizing EEG signal inputs and deep learning algorithms is presented in Figure 2. The procedure has four major phases: Data Collection, Data Preprocessing, Deep Learning Models, and Output Classification. Data collection refers to the methodology employed in acquiring the necessary brain wave data for forecasting concentration levels using an EEG device, the parameters included in the dataset, the number of features present, and the associated concentration levels for each recorded instance. The subsequent phase is data preprocessing, which involves preparing the data as input for deep learning algorithms. The process entails cleaning the data to eliminate redundant data, correlation analysis to identify pertinent features for predicting their target variable, feature scaling to prevent bias in weight assignment during deep learning model training, and one-hot encoding for the transformation of the target variable. The subsequent phase involves training various deep learning models—ANN, LSTM, and GRU—and evaluating their accuracies on test data to assess their performance on unknown data. The data input for these algorithms is categorised into four distinct classes: high concentration, medium-high concentration, medium-low concentration, and low concentration.

Table 2: Concentration Level for Corresponding Tasks.

Task	Concentration Level
Performing Arithmetic Calculations	High
Reading Technical Articles, listening to technical podcasts	Medium-High
Reading the transcripts	Medium-High
Browsing the Internet	Medium-Low
Just sitting there, eyes open or closed	Low

Source: Authors (2026).

III.1 DATA COLLECTION

The dataset used for predicting an individual's concentration level during activities of varying levels of attention is shown in Figure 3. An individual wears the Muse EEG device on their scalp while participating in five activities. Table 2 gives the activities' details and associated concentration levels. The EEG device has four electrodes, to which we assign the numbers 0, 1, 2, and 3. Five different types of brain waves—delta, theta, alpha, beta, and gamma—are captured from each electrode while performing each task.

The captured brain waves are labelled with their names and subscripted with the electrode number to neatly identify the features in the dataset. For instance, if electrode 0 captures waves in the frequency range of 0.5 to 4Hz (delta wave frequency), we will refer to that frequency as "delta0". In this way, all 20 features (five types of brain waves from each of the four electrodes) in the dataset are captured by an EEG device and labelled according to the given method. For each record, the concentration level (low, medium low, medium high, and high) has been assigned as shown in Table 2. This dataset serves as the basis for identifying an individual's cognitive engagement.

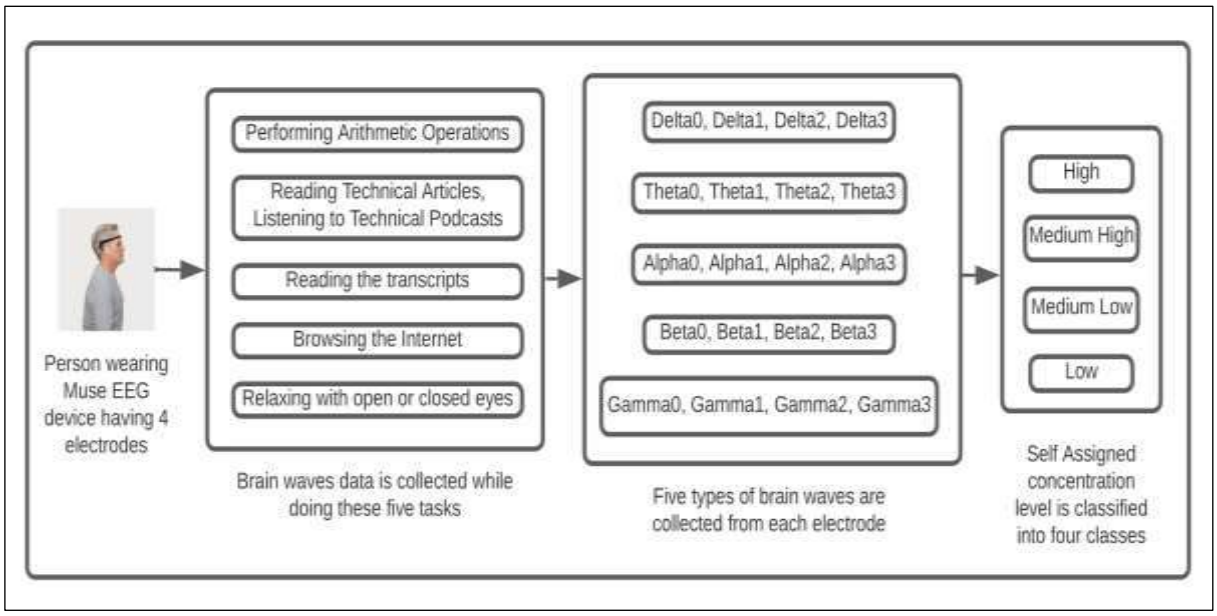


Figure 3: Data Collection Process.
Source: Authors, (2026).

III.2 DATA PREPROCESSING

Data preprocessing is the most crucial step after data collection and before applying any deep learning algorithm. The following four steps are performed to prepare the collected brain wave data in the required format to provide as input to the algorithms.

a) **Data Cleaning:** When an EEG device's electrodes capture brain wave data, there is a strong possibility that eye movements or even slight head movements could introduce noise into the data. Eliminating that noise is important before going ahead to ensure the correct predictions. The integrated software on the device has already removed noise from the collected dataset.

b) **Correlation Analysis:** Correlation analysis involves finding the direct relationship between the target variable and the input features of the dataset. There are two types of correlation: positive correlation, which means if input features are increasing, then the value of the target variable will also increase, and vice-versa. Another type is negative correlation, which means if input features are increasing, then the value of the target variable will decrease, and vice versa. In the current study, the target variable is an individual's concentration level, while the input features are five types of brain waves captured from four electrodes. Correlation analysis proves no positive or negative correlation between the input features and target variables.

c) **Feature Scaling:** Due to the varied frequency ranges in brain wave data, deep learning models might give greater importance to higher-magnitude values during training and less importance to lower-magnitude ones. To address this possible bias, a scaling approach is used to normalize the data by adjusting each feature's minimum and maximum values to 0 and 1, respectively. The following formula will calculate the new scaled value:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

This formula normalizes the values of all features. This process ensures that all features are treated equally, preventing magnitude-based bias in the training process.

d) **One Hot Encoding:** The concentration level feature is converted using one-hot encoding, a technique that transforms categorical data into numerical arrays with a length equal to the number of unique categories as stated by [12]. For this case, we have four distinct categories of concentration levels, i.e., High concentration, Medium high concentration, Medium low concentration, and Low concentration, resulting in an array length of four.

This procedure produces binary vectors, in which the element matching the appropriate class is assigned a value of one while all other elements are set to zero. For example, if we had a brain wave record that indicates high concentration during a task, the corresponding one-hot encoded array would be [1 0 0 0], where 1 will denote the current record is for the High concentration level task, and other classes will be set to 0. In this way, brain wave records are encoded using one-hot encoded arrays and then used as input for future processing and analysis.

III.3 DEEP LEARNING ALGORITHMS

Preprocessed data is provided as input to three deep learning algorithms: Artificial Neural Network (ANN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). All the models use one input layer, three hidden layers, and one output layer. The output layer contains four neurons and uses the softmax activation function, which is best for multiclass classification problems. All three models are trained with 70% and tested on 30% of the data. A categorical cross-entropy loss function is used since predicting the concentration level is a multi-class classification problem. To update the weights while training these models, the Adam optimizer is used as an optimization algorithm instead of the classical stochastic gradient descent procedure.

a) **Artificial Neural Network (ANN):** The neuron is the fundamental component of an Artificial Neural Network (ANN). ANN consists of several linked neurons that process information by applying weights and activation functions. Each time the neuron receives an input, a random weight is assigned to that value, which shows the significance of that value in the prediction process. Subsequently, all the data is combined and given to an activation function. This activation function introduces non-linearity in the data. Lastly, the output is computed. The error is backpropagated, and the weights are modified based on the difference between the calculated and actual output to minimize the difference. The accuracy of an ANN increases as the difference decreases. The preprocessed data of the brain waves is given as input to the ANN to predict the concentration level while doing different cognitive tasks. The architecture of the ANN model is illustrated in Figure 4. The input layer of the ANN consists of 12 neurons and employs the ReLU activation function. It features three hidden layers with 128, 256, and 128 neurons, each utilizing the ReLU activation function. The output layer comprises four neurons, which classify the data into four different concentration classes and use the softmax activation function, which is optimal for multiclass classification tasks.

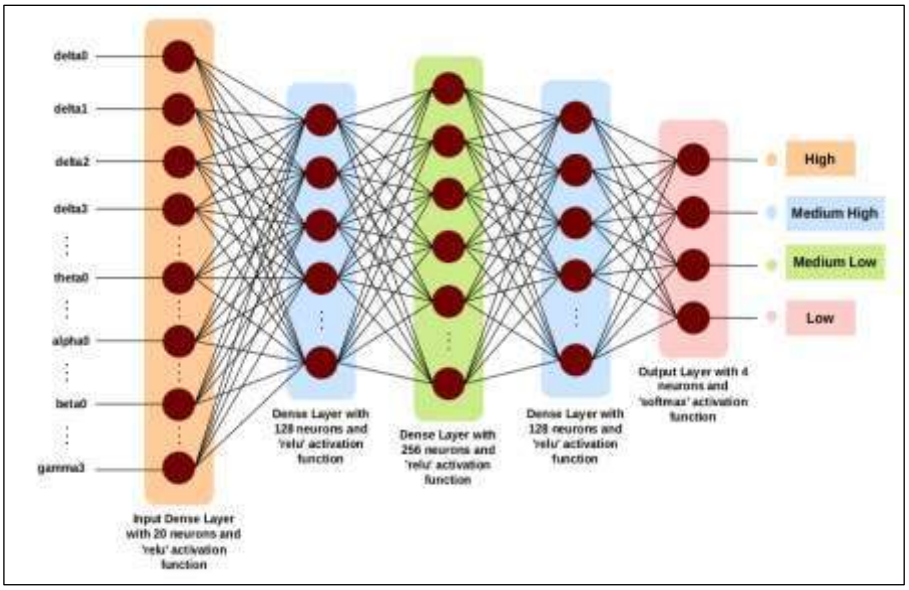


Figure 4: Architecture of Proposed Artificial Neural Network.
Source: Authors, (2026).

b) **Long Short Term Memory (LSTM):** Brain wave data refers to the frequency of electrical activity recorded over time, similar to time series data. Recurrent neural networks do well in processing time series data, but cannot learn long-term dependencies. LSTM networks are a specific form of recurrent neural networks that successfully overcome this drawback. The fundamental structure of LSTM includes the cell state, the input gate, the output gate, and the forget gate. The cell state is the central part of the LSTM network, which stores long-term dependencies and decides which information to let go of and which information to send for further analysis. The other three gates regulate the flow of information: the input gate admits new data, the output gate outputs the information, and the forget gate erases redundant data. The LSTM network uses the truncated backpropagation of error to enhance its accuracy and train it to understand long-term data dependencies. The architecture of the LSTM has been illustrated in Figure 5.
It consists of three LSTM layers, each containing 128, 256, and 128 neurons. The activation function used in these layers is the Rectified Linear Unit (ReLU). The input comprises preprocessed and reshaped data that is suitable for LSTM cells. After undergoing additional processing through three LSTM layers, the data ultimately reaches the output layer, which consists of four neurons for each concentration level class. The softmax activation function is employed as it is the most suitable choice for multiclass classification tasks.

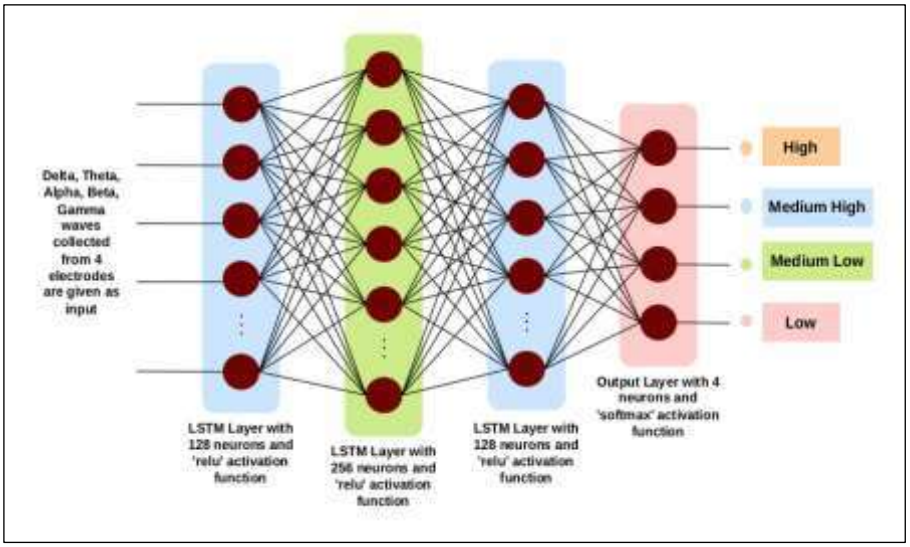


Figure 5: Architecture of Proposed Long Short-Term Memory.
Source: Authors, (2026).

c) **Gated Recurrent Unit (GRU):** GRUs is a kind of RNN that addresses the limitations of LSTM networks and traditional RNNs. Unlike LSTM, which is a cell state, the basic unit of GRU is the hidden state. Compared to LSTM, the GRU has fewer gates. The gates include the reset gate, which forgets redundant information, and the update gate, which balances past information and new important data for longer. Due to its reduced number of gates and parameters, the GRU model can be trained more quickly than the LSTM. The architecture of the GRU model has been illustrated in Figure 6. It consists of one input layer, three dense GRU layers, each containing 128, 256, and 128 neurons, and the output layer, which consists of four neurons for each concentration level class. The ReLU is the activation function in these layers. The input comprises preprocessed and reshaped data that is suitable for GRU cells. After undergoing additional processing through three GRU layers, the data ultimately reaches the output layer, where the softmax activation function is employed because it is the most suitable choice for multiclass classification tasks.

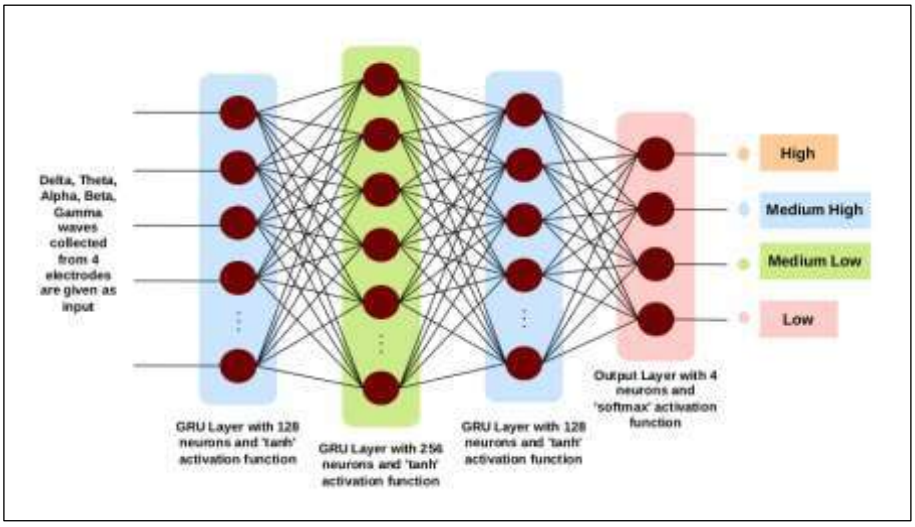


Figure 6: Architecture of Proposed Gated Recurrent Unit.
Source: Authors, (2026).

IV. PERFORMANCE MEASURES

Performance evaluation ensures the proposed methodology's efficiency, validity, and applicability. The proposed models are evaluated using the following metrics.

a) **Confusion Matrix:** A confusion matrix is a crucial performance evaluation tool that encapsulates the accuracy of a classification model. It helps to determine where our model underperforms when differentiating between classes. The matrix enumerates the counts of true positives (TP), which reveals how frequently a model accurately identifies a positive sample as positive; true negatives (TN), which means the frequency with which a model labels a positive sample as negative; false positives (FP), which indicate how frequently a model wrongly labels a negative sample as positive; and false negatives (FN), which reflects how often a model accurately labels a positive sample as Negative. According to data analysis, these four categories can be specified for the high concentration class as follows: TP refers to the number of instances in which the recorded brain waves indicative of high concentration were accurately classified by proposed algorithms.

TN relates to cases where the proposed algorithms accurately predict brain wave recordings classified as non-high concentrations belonging to other categories, such as medium-high, medium-low, or low concentrations. FP denotes false positives, indicating the frequency of brain wave records that belong to different classes but are incorrectly classified by the proposed system as belonging to the high-concentration class. FN denotes false negatives, indicating the frequency of brain wave recordings associated with high-concentration tasks misclassified as belonging to concentration classes other than high-concentration activities when processed by the proposed system. Specifically, a confusion matrix is an N × N matrix employed to assess the performance of a classification model, where N denotes the total number of target classes. It compares the actual values of the target class in the dataset with the predicted target class by the proposed model; if both match, then the prediction is correct; otherwise, it is not. It helps to identify how well our classification model performs and what kinds of errors it makes.

b) **Accuracy**: Accuracy metrics provide insight into how well the proposed model classifies a specific data set into the appropriate class. Accuracy is calculated as the percentage of accurate predictions over all other predictions.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{2}$$

Equation 2 provides the formula for calculating accuracy and finds the per centage of correctly classified brain wave records to each class among all the predictions.

c) **Precision:** Precision measures the percentage of true predictions within all positive predictions. The proportion of accurately categorized positive samples (TP) to the total number of positively classified samples (either correctly or incorrectly) is known as precision, as given in Equation 3.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

Regarding the classification problem, precision is the rate at which the proposed model accurately classifies a given brain wave record to the actual concentration class. The precision value increases as the number of FP decreases. FP refers to the number of records that do not belong to a specific class but are falsely classified as belonging to that class. FP increases when the model has not accurately learned the parameters to detect records that do not belong to the current class. That is why the model erroneously classifies the brain record as positive. Increased precision indicates that the model has effectively learned the parameters that allow it to classify records accurately, distinguishing which belong to a specific class.

d) **Recall:** The recall is determined as the proportion of correctly identified positive samples, i.e., TP to the total number of actual positive samples, which encompasses both TP and those that are positive but incorrectly identified as negative (FN), as given in Equation 4 as shown below:

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4}$$

The recall value increases as the FN decreases. FN represents the count of records that belong to a specific class but are misclassified as belonging to another class. FN will increase when the model has not accurately learned the parameters necessary to identify data belonging to the current class. That is why the model erroneously categorises the brain record as negative. A higher recall indicates the model has not effectively learned the parameters to classify this data into the right class.

e) **F1 Score:** The F1 Score, defined as the harmonic mean of precision and recall, offers a consolidated metric that integrates precision and recall, thus equilibrating the trade-off between these two measures. The formula to calculate the F1 Score using precision and recall is given in Equation 5.

$$F1\ Score = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{5}$$

Since FP influences precision, and FN influences recall. Evaluating both precision and recall when computing the F1 Score enables one to assess the model's capacity to identify FP and FN. As the number of data points for each class differs in an unbalanced dataset, the proposed model must discern the parameters that accurately classify each record as belonging or not belonging to a specific class. If precision is the sole performance indicator, the model's capacity to accurately identify positive records may be overlooked. On the other hand, focusing solely on recall could hinder the model's capacity to identify bad records. Both features are essential for an effective model. The F1 Score is an excellent indicator for assessing model performance, as it mitigates the substantial bias inherent in precision and recall.

V. RESULT ANALYSIS AND DISCUSSION

This section presents the results of implementing deep learning methods for predicting concentration levels, demonstrating how an individual engages cognitively in different attention activities. The dataset has 20 features, including five unique brain wave frequencies, each recorded from four electrodes. The data set undergoes data preprocessing, which includes cleaning, correlation analysis, feature scaling, and one-hot encoding of the target variable. After this, the dataset is fed into three deep-learning algorithms to predict concentration levels in four categories. The preprocessed dataset is partitioned into two subsets, with 70% of the data utilized for training the model and the remaining 30% used for evaluating its performance.

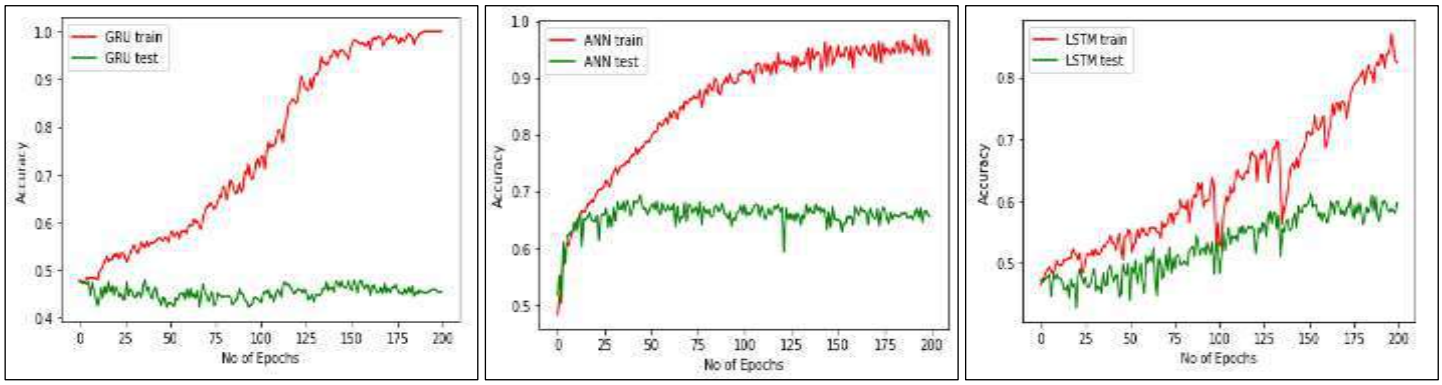
The models, including ANN, LSTM, and GRU, are trained for 200 epochs using different batch sizes. The ANN model uses a batch size of 5, the LSTM model uses a batch size of 64, and the GRU model uses a batch size of 32. The additional parameters for training these models are specified in Table 3. The performance of models is evaluated using various measures, including accuracy, precision, recall, and F1 Score.

Table 3: Parameter values used to train deep learning models.

Parameter	Values
Optimizer	Adam
Loss Function	Categorical Cross-entropy
Metric	Accuracy
Epoch	200

Source: Authors (2026).

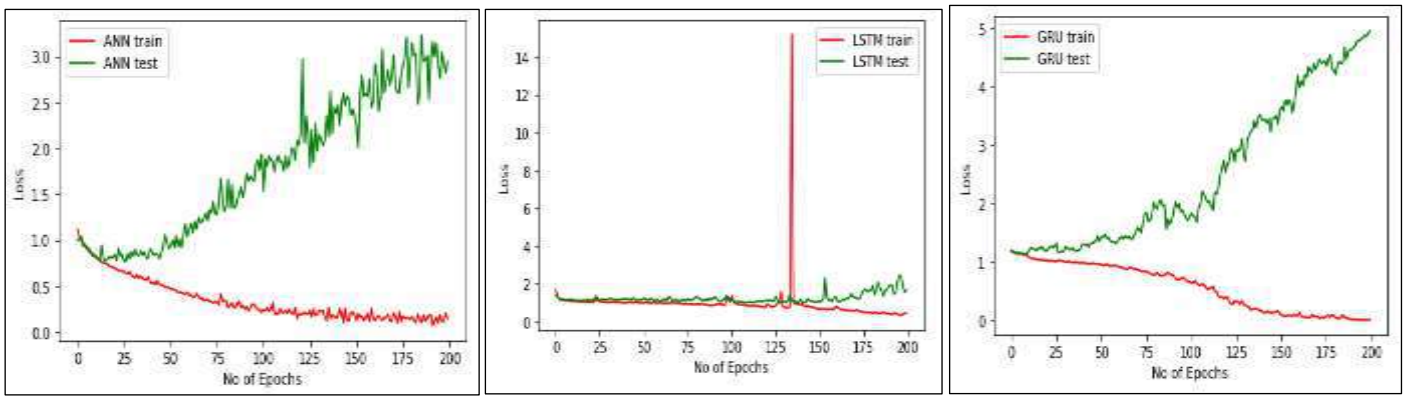
Figure 7 shows that ANN's training and testing accuracies steadily increase up to 50 epochs. After that, although there is an increase in training accuracy, the testing accuracy shows only a slight improvement. In the case of LSTM, both the training and testing accuracies show fluctuations until the 125th epoch. However, after this point, the testing accuracy remains constant, regardless of the continued increase in training accuracy. The GRU model consistently improves training accuracy up to 100 epochs, whereas testing accuracy remains relatively stable. The ANN, LSTM, and GRU models demonstrate overfitting of the data at epochs 50, 125, and 50, respectively.



(a) GRU (b) ANN (c) LSTM

Figure 7: Training and Testing Accuracy of Deep Learning Algorithms.
Source: Authors (2026).

The same phenomena are observed from the loss graphs in Figure 8. The training and testing loss for ANN reduces up to 50 epochs, LSTM up to 125 epochs, and GRU up to 100 epochs. However, after that point, only the training loss decreases while the testing loss increases, suggesting that the models are overfitting the data. Figure 9 compares the classification accuracy of algorithms ANN, LSTM, and GRU. The ANN achieves maximum accuracy, which is 65.63%, followed by LSTM, which is 59.67%, and GRU networks, which is 45.39%.



(a) ANN (b) LSTM (c) GRU

Figure 8: Training and Testing Loss of Deep Learning Algorithms.
Source: Authors (2026).

The imbalance of the data set, as indicated in Table 4, compromises the reliability of accuracy as a sole metric for assessing the performance of deep learning models. There is a need to determine the model's ability to correctly classify the major class (containing a more significant number of records) and the minor class (containing a smaller number of records). The accuracy of the models for each class is provided in the confusion matrix for each deep learning technique mentioned in Table 5. The confusion matrix indicates that the ANN model exhibits a higher count of TP and TN than the LSTM and GRU models. Precision verifies the model's quality, determining the accurate classification of positive samples from the total number of positive samples.

As depicted in Figure 10a, the ANN model exhibits marginally superior precision to the LSTM network for the low and medium high concentration classes. Conversely, the LSTM network demonstrates higher precision than the ANN for the high and medium low concentration classes. Recall is a measure of quantity that identifies the number of correctly classified positive samples from the total number of positive samples. A model with a high recall value will yield more relevant results. Based on the data presented in Figure 10b, it is evident that ANN exhibits superior performance in tasks involving high and medium-high concentrations. On the other hand, LSTM networks demonstrate better performance in tasks involving low concentrations. ANN and LSTM networks exhibit similar recall values for tasks involving medium-low concentrations.

Table 4: Number of records for each concentration level class.

Class	No. of Records
High Concentration	3410
Medium_High Concentration	501
Medium_Low Concentration	909
Low Concentration	2423

Source: Authors, (2026)

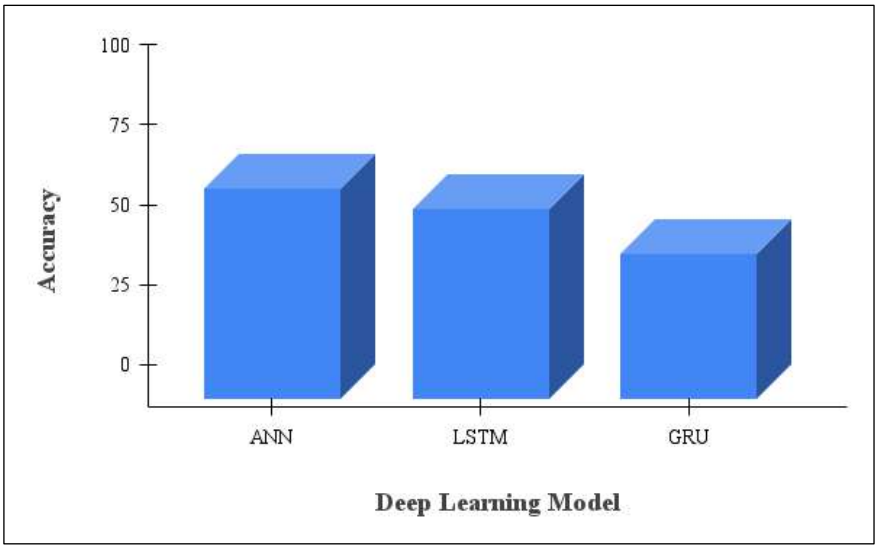


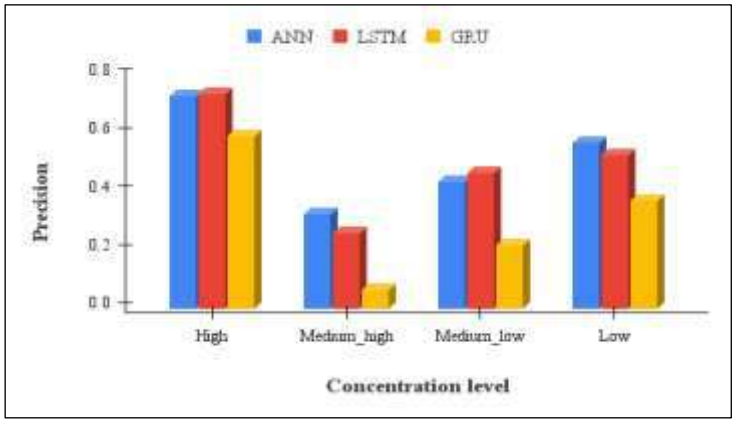
Figure 9: Comparison of Accuracies of Deep Learning Algorithms.
Source: Authors, (2026).

Support is a measure of the class frequency that appears in the dataset. Due to the imbalance in the dataset, which can be seen in Figure 10d, the precision and recall values are likely to be biased. To address this issue, the F1 Score metric, the harmonic mean of precision and recall, is computed to measure the model's overall performance. Figure 10c shows that the ANN model has a higher F1 Score for high and medium high concentrations classes than other models. For low and medium low concentration classes, LSTM networks have a slightly higher F1 Score than ANN.

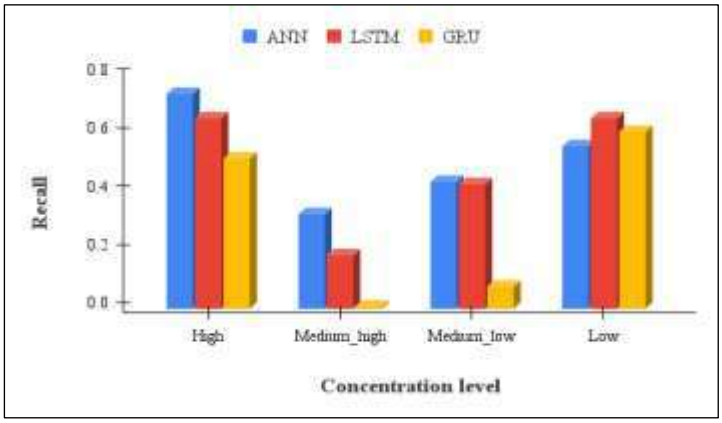
Table 5: Confusion Matrix of Deep Learning Algorithms.

Algorithm	True Value	Predicted Value			
		High	Medium_High	Medium_Low	Low
ANN	High	575	38	27	74
	Medium_High	36	32	8	19
	Medium_Low	32	13	56	77
	Low	117	11	46	288
LSTM	High	547	34	44	209
	Medium_High	53	22	10	34
	Medium_Low	27	10	100	96
	Low	114	18	60	380
GRU	High	430	6	45	353
	Medium_High	44	1	9	65
	Medium_Low	42	2	18	171
	Low	207	6	10	349

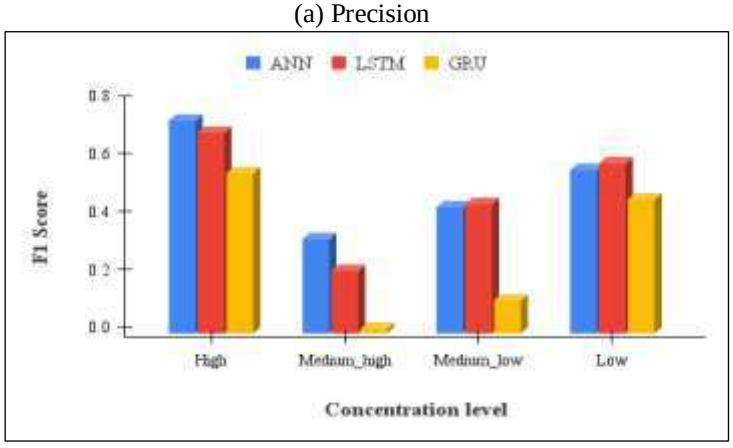
Source: Authors, (2026).



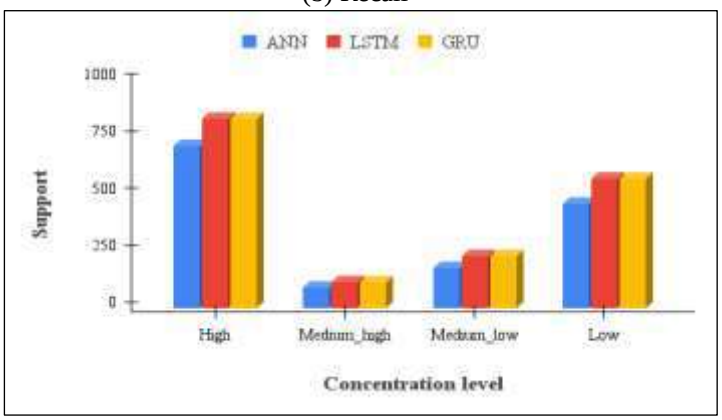
(a) Precision



(b) Recall



(c) F1 Score



(d) Support

Figure 10: Performance measures of Deep Learning Algorithms.
Source: Authors, (2026).

VI. CONCLUSION

This study used recorded brain signals from the Muse EEG device to compare the effectiveness of three deep learning algorithms in predicting concentration levels while performing various cognitive tasks such as arithmetic calculations, reading technical articles, listening to technical podcasts, reading transcripts, browsing the Internet, and simply relaxing with closed or open eyes. Deep learning algorithms, such as ANN, LSTM, and GRU, are used to analyse the brain signals; of these, ANN provides the best accuracy (65.63%), followed by LSTM (59.67%) and GRU (45.39%). ANN outperformed LSTM and GRU networks in accurately predicting concentration levels by analysing EEG data.

When evaluating the F1 Score, ANN demonstrated superior performance in classes with high and medium-high concentrations. Still, LSTM networks perform better for classes with low and medium-low concentrations than the other two networks. This study utilizes the brainwave data obtained from the Muse EEG equipment, equipped with four electrodes, to predict the degree of concentration. To further enhance this research in the future, one can augment the data collection process by collecting brainwave data from several people using an EEG device equipped with many channels. Subsequently, the data collected can be examined to identify the underlying trends in the brainwave data and to build a reliable deep-learning model to evaluate the students' cognitive involvement.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Yogesh Patil, Snehal Mali, and Harsha Gaikwad.
Methodology: Harsha Gaikwad, Manjushree Laddha, and Arvind Kiwelekar.
Investigation: Yogesh Patil, Snehal Mali.
Discussion of results: Yogesh Patil, Snehal Mali, and Harsha Gaikwad.
Writing – Original Draft: Yogesh Patil, Snehal Mali, and Harsha Gaikwad.
Writing – Review and Editing: Sanil Gandhi and Harsha Gaikwad.
Resources: Harsha Gaikwad, Manjushree Laddha, Yogesh Patil, and Snehal Mali.
Supervision: Manjushree Laddha, and Arvind Kiwelekar.
Approval of the final text: Yogesh Patil, Snehal Mali, Harsha Gaikwad, Sanil Gandhi, Manjushree Laddha, and Arvind Kiwelekar.

VIII. REFERENCES

[1] Q. He, O. Turel, and A. Bechara, “Brain anatomy alterations associated with social networking site (sns) addiction,” Scientific Reports, vol. 7, no. 1, p. 45064, Mar 2017.

[2] H. Gardner, The Disciplined Mind: What All Students Should Understand. New York: Simon & Schuster, 1999.

[3] M. A. A. Dewan, M. Murshed, and F. Lin, “Engagement detection in online learning: a review,” Smart Learning Environments, vol. 6, no. 1, p. 1, Jan 2019.

[4] M. Piwowarski, U. S. Singh, and K. Nermend, “Application of eeg metrics in the decision-making process,” in Experimental and Quantitative Methods in Contemporary Economics, K. Nermend and M. Latuszy´nska, Eds. Cham: Springer International Publishing, 2020, pp. 187–199.

[5] Aayesha, M. B. Qureshi, M. Afzaal, M. S. Qureshi, and M. Fayaz, “Machine learning-based eeg signals classification model for epileptic seizure detection,” Multimedia Tools and Applications, vol. 80, no. 12, pp. 17 849–17 877, 2021..

[6] A. M. Maitin, J. P. Romero Mu˜noz, and ´A. J. Garc´ıa-Tejedor, “Survey of machine learning techniques in the analysis of eeg signals for parkinson’s disease: A systematic review,” Applied Sciences, vol. 12, no. 14, p. 6967, 2022.

[7] D. Pirrone, E. Weitschek, P. Di Paolo, S. De Salvo, and M. C. De Cola, “Eeg signal processing and supervised machine learning to early diagnose alzheimer’s disease,” Applied sciences, vol. 12, no. 11, p. 5413, 2022.

[8] T. Sudhakar, G. H. Krishnan, N. Krishnamoorthy, B. Janney, M. Pradeepa, and J. Raghavi, “Sleep disorder diagnosis using eeg based deep learning techniques,” in 2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII). IEEE, 2021, pp. 1–4.

[9] R. Alhalaseh and S. Alasasfeh, “Machine-learning-based emotion recognition system using eeg signals,” Computers, vol. 9, no. 4, p. 95, 2020.

[10] J. J. Bird, A. Ekart, C. D. Buckingham, and D. R. Faria, “Mental emotional sentiment classification with an eeg-based brain-machine interface,” in Proceedings of theInternational Conference on Digital Image and Signal Processing (DISP’19), 2019.

[11] H. Gaikwad, S. Gandhi, A. Kiwelekar, and M. Laddha, “Analyzing brain signals for predicting students’ understanding of online learning: A machine learning approach,” International Journal of Performability Engineering, vol. 19, no. 7, p. 462, 2023.

[12] V. T. Lokare and L. D. Netak, “Concentration level prediction system for the students based on physiological measures using the eeg device,” in Intelligent Human Computer Interaction: 12th International Conference, IHCI 2020, Daegu, South Korea, November 24–26, 2020, Proceedings, Part I 12. Springer, 2021, pp. 24–33.

[13] A. Al-Nafjan and M. Aldayel, “Predict students’ attention in online learning using eeg data,” Sustainability, vol. 14, no. 11, p. 6553, 2022.

[14] N. Friedman, T. Fekete, K. Gal, and O. Shriki, “Eeg-based prediction of cognitive load in intelligence tests,” Frontiers in human neuroscience, vol. 13, p. 191, 2019.

[15] A. Farkish, A. Bosaghzadeh, S. H. Amiri, and R. Ebrahimpour, “Evaluating the effects of educational multimedia design principles on cognitive load using eeg signal analysis,” Education and Information Technologies, vol. 28, no. 3, pp. 2827–2843, 2023.

[16] X. Liu, P.-N. Tan, L. Liu, and S. J. Simske, “Automated classification of eeg signals for predicting students’ cognitive state during learning,” in Proceedings of the International Conference on Web Intelligence, 2017, pp. 442–450.

[17] N. Jamil and A. N. Belkacem, “Advancing real-time remote learning: A novel paradigm for cognitive enhancement using eeg and eye-tracking analytics,” IEEE Access, vol. 12, pp. 93 116–93 132, 2024.

[18] F. J. Ram´ırez-Arias, E. E. Garc´ıa-Guerrero, E. Tlelo-Cuautle, J. M. Colores-Vargas, E. Garc´ıa-Canseco, O. R. L´opez-Bonilla, G. M. Galindo-Aldana, and E. Inzunza Gonz´alez, “Evaluation of machine learning algorithms for classification of eeg signals,” Technologies, vol. 10, no. 4, p. 79, 2022.

[19] S. K. Gatfan, “A review on deep learning for electroencephalogram signal classification,” Journal of Al-Qadisiyah for Computer Science and Mathematics, vol. 16, no. 1, pp. 137–151, 2024.

[20] E.-A. Paraschiv, L. B˜ajenaru, C. Petrache, O. Bica, and D.-N. Nicolau, “Ai-driven neuro-monitoring: Advancing schizophrenia detection and management through deep learning and eeg analysis,” Future Internet, vol. 16, no. 11, 2024.

[21] B. Siregar, G. Florence, S. Seniman, F. Fahmi, and N. Mubarakah, “Classification of human concentration levels based on electroencephalography signals,” JOIV: International Journal on Informatics Visualization, vol. 8, no. 2, pp. 923–930, 2024.

[22] A. Azhari and F. I. Ammatulloh, “Classification of concentration levels in adult-early phase using brainwave signals by applying k-nearest neighbor,” Signal and Image Processing Letters, vol. 1, no. 1, pp. 14–24, 2019.

[23] H. Zeng, C. Yang, G. Dai, F. Qin, J. Zhang, and W. Kong, “Eeg classification of driver mental states by deep learning,” Cognitive neurodynamics, vol. 12, pp. 597–606, 2018.

[24] J. Slater, R. Joober, B. L. Koborsy, S. Mitchell, E. Sahlas, and C. Palmer, “Can electroencephalography (eeg) identify adhd subtypes? a systematic review,” Neuroscience Biobehavioral Reviews, vol. 139, p. 104752, 2022.

[25] A. Rehman, I. Heldal, and J. C.-W. Lin, “Ssrepl-adhd: Adaptive complex representation learning framework for adhd detection from visual attention tasks,” in 2024 IEEE International Conference on Big Data (BigData). IEEE, 2024, pp. 8164–8172.

[26] G. Cisotto, A. Zanga, J. Chlebus, I. Zoppis, S. Manzoni, and U. Markowska-Kaczmar, “Comparison of attention-based deep learning models for eeg classification,” arXiv preprint arXiv:2012.01074, 2020.

[27] C¸ . ˙I. Acı, M. Kaya, and Y. Mishchenko, “Distinguishing mental attention states of humans via an eeg-based passive bci using machine learning methods,” Expert Systems with Applications, vol. 134, pp. 153–166, 2019.

[28] N. S. Williams, G. M. McArthur, and N. A. Badcock, “10 years of epoc: A scoping review of emotiv’s portable eeg device,” BioRxiv, 2020.

[29] L. E. Ismail and W. Karwowski, “Applications of eeg indices for the quantification of human cognitive performance: A systematic review and bibliometric analysis,” PLOS ONE, vol. 15, no. 12, pp. 1–33, 12 2020.

[30] S. D. You, “Classification of relaxation and concentration mental states with eeg,” Information, vol. 12, no. 5, p. 187, 2021.

[31] D. C. E. Saputra, A. Azhari, and A. Ma’arif, “K-nearest neighbor of beta signal brainwave to accelerate detection of concentration on student learning outcomes.” Engineering Letters, vol. 30, no. 1, 2022.

[32] A. U. Rehman, X. Shi, F. Ullah, Z. Wang, and C. Ma, “Measuring student attention based on eeg brain signals using deep reinforcement learning,” Expert Systems with Applications, vol. 269, p. 126426, 2025.

[33] M. S. Dehnavi, V. S. Dehnavi, and M. Shafiee, “Classification of mental states of human concentration based on eeg signal,” in 2021 12th International Conference on Information and Knowledge Technology (IKT). IEEE, 2021, pp. 78–82.

[34] S. Basterrech and P. Kr¨omer, “A nature-inspired biomarker for mental concentration using a single-channel eeg,” Neural Computing and Applications, vol. 32, no. 12, pp. 7941–7956, 2020.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 684-695. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2990

RESEARCH ARTICLE OPEN ACCESS

BREAST CANCER DETECTION USING LATE FUSION APPROACH ON MULTIMODAL DATA WITH DEEP LEARNING

Nathan Daud[1], Firhan Imam Haekal[2], Adinda Fatimah Az-Zahra[3], and Fitri Utaminingrum[4]

[1,2,3,4]Department of Informatics Engineering, Brawijaya University, Malang, Indonesia.

[1]https://orcid.org/0009-0004-2141-4138, [2]https://orcid.org/0009-0006-6622-9926
[3]https://orcid.org/0009-0004-1821-8961, [4]https://orcid.org/0000-0002-0281-9429

Email: nathandaud@student.ub.ac.id, firhanih@student.ub.ac.id, azzahraadinda12@student.ub.ac.id, firhanih@student.ub.ac.id

ARTICLE INFO

Article History
Received: November 22, 2025
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Breast cancer,
Deep learning,
Late fusion,
Probability multiplication,
Medical imaging.

ABSTRACT

Breast cancer stands as one of the most formidable challenges in contemporary global healthcare, with approximately 18.1 million new cancer cases reported worldwide and an urgent need for revolutionary advances in early detection methodologies. Current breast cancer diagnostic methodologies present significant limitations through unimodal approaches, where mammography suffers from masking effects and false positives/negatives, thermography demonstrates low specificity of 57.8%, and clinical tabular data alone proves insufficient for definitive diagnosis. This research presents a comparative study of four late fusion strategies for breast cancer diagnosis, integrating predictions from deep learning models trained on mammography images, thermography images, and clinical tabular data. The fusion methods evaluated include probability multiplication (product rule), weighted averaging, stacking metaclassifier, and log-opinion pool fusion. The log-opinion pool fusion method achieved superior performance with 97.5% overall accuracy, surpassing the other fusion approaches and all individual. The method has also achieved precision of 0.91–0.95 for benign cases 1.00 for malignant cases, with fusion approaches maintaining zero false positive rate and recall for malignant cases up to 95%.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Breast cancer stands as one of the most formidable challenges in contemporary global healthcare. It has consistently maintained its position as the most frequently diagnosed cancer among women globally, while simultaneously serving as a leading cause of cancer-related mortality. The World Health Organization's data reveal approximately 18.1 million new cancer cases reported worldwide in 2020, where breast cancer contributed significantly to this staggering figure, ultimately resulting in 9.6 million cancer-related deaths [1], hence the urgent and critical need for revolutionary advances in early detection methodologies, as the window of opportunity for successful treatment narrows dramatically with delayed diagnosis.

Medical evidence consistently demonstrates that patients diagnosed during the early stages of breast cancer experience dramatically improved treatment outcomes, enhanced therapeutic efficacy, and substantially higher survival rates compared to those diagnosed at advanced stages [2]. However, the current landscape of breast cancer diagnostic methodologies presents significant limitations that compromise the reliability and accuracy of early detection efforts. This diagnostic inadequacy represents a critical gap in healthcare delivery that demands immediate attention and innovative solutions.

In parallel to diagnostic innovation, there is growing interest in novel, non-thermal therapeutic modalities for breast cancer. For instance, plasma-activated water has demonstrated selective cytotoxic effects against breast cancer cells, representing a promising direction for integrated therapeutic strategies alongside advances in diagnostic imaging [3].Contemporary breast cancer detection predominantly relies on unimodal diagnostic approaches, each carrying inherent limitations that collectively contribute to suboptimal diagnostic accuracy [4].

Mammography faces substantial challenges that compromise its diagnostic reliability such as dense fibroglandular tissue that can effectively mask tumors, creating a phenomenon known as masking effect, which substantially reduces the visibility of malignant lesions [5]. Furthermore, mammography demonstrates concerning vulnerability to both false negative results, where existing cancers remain undetected, and false positive results, which incorrectly indicate the presence of cancer. Advanced segmentation methods such as the Expectation-Maximization Gaussian Mixture Model (EM-GMM) have shown promise for improving the identification of suspicious areas in mammograms, supporting subsequent analysis and diagnosis [6]. False negatives delay crucial treatment interventions, while false positives generate unnecessary patient anxiety and often trigger invasive follow-up procedures that could have been avoided with more accurate initial screening.

Thermography, while offering promising advantages as a non-invasive and radiation-free diagnostic modality that detects metabolic activity-related temperature changes associated with tumor presence, presents its own set of critical limitations. Historical analysis reveals that thermography suffers from significantly lower specificity compared to mammography, with documented specificity rates of only 57.8% compared to mammography's 73.3% [7]. This reduced specificity stems primarily from thermography's inability to detect minute calcium deposits, known as microcalcifications, which are often successfully identified through mammographic examination. The poor specificity rate of 55.7% attributed to high False Positive Rates (FPR) further undermines thermography's reliability as a standalone diagnostic tool.

Clinical tabular data, exemplified by datasets such as the Breast Cancer Wisconsin (Diagnostic) dataset, provides valuable numerical insights derived from fine needle aspiration (FNA) analysis, including cellular characteristics such as size, shape, and texture. However, when utilized in isolation, this data modality proves insufficient for definitive diagnosis. While these numerical features originate from image analysis and offer detailed cellular-level information, they fundamentally lack the comprehensive visual representation of tissues and cells that imaging modalities provide. The emergence of multimodal diagnostic approaches represents a paradigm shift in breast cancer detection methodology, offering unprecedented potential to overcome the limitations inherent in unimodal systems.

This performance is comparable to other healthcare applications where binary classification methods have achieved high accuracy with fast computational performance [8]. By strategically combining information from multiple data sources, multimodal approaches can leverage complementary information streams to achieve more accurate and comprehensive diagnostic outcomes. However, despite the considerable research efforts invested in this domain, significant opportunities remain for advancement, particularly in the development of optimal fusion strategies that can maximize the diagnostic potential of multimodal data integration. This research presents a novel late fusion approach for breast cancer diagnosis that uses probability multiplication to combine predictions from independent deep learning models trained on mammography images, thermography images, and clinical tabular data.

Unlike traditional fusion methods that rely on simple averaging, this sophisticated probability multiplication strategy strengthens diagnostic confidence when models agree while maintaining robustness during disagreement, preserving each modality's independent feature extraction while enabling advanced decision-level integration. The methodology addresses critical gaps in current multimodal medical diagnostics by developing specialized architectures for each data type, with the final classification determined by the highest probability after mathematical multiplication of individual model outputs. This timely research leverages the growing availability of diverse medical data and advancing deep learning technologies to potentially improve diagnostic accuracy, reduce healthcare costs associated with false positives and negatives, minimize patient anxiety, and ultimately save lives through earlier and more reliable breast cancer detection.

II. MATERIALS AND METHODS

This research implements an end-to-end multimodal late-fusion framework for breast cancer detection that independently processes mammography, thermography, and structured clinical data through modality-specialized networks (ResNet50 for imaging an a DNN for tabular data), producing class-posterior probabilities per branch for decision-level integration, as illustrated in . Four alternative late-fusion strategies are comparatively assessed at inference, that is product rule, weighted averaging, log-opinion pool, and stacking meta-classifier while preserving modular feature extraction and auditability for each modality. The solution is evaluated based on accuracy, precision, recall, F1-score, macro/weighted averages, and confusion matrices metrics with a balanced 40-sample test set across all branches and fusion rules. The final model is selected by optimizing a predefined validation objective (e.g., macro-F1 or accuracy) under clinically motivated sensitivity–specificity considerations, and the chosen operating point is documented with per-class metrics and confusion matrices for interpretability and clinical relevance.

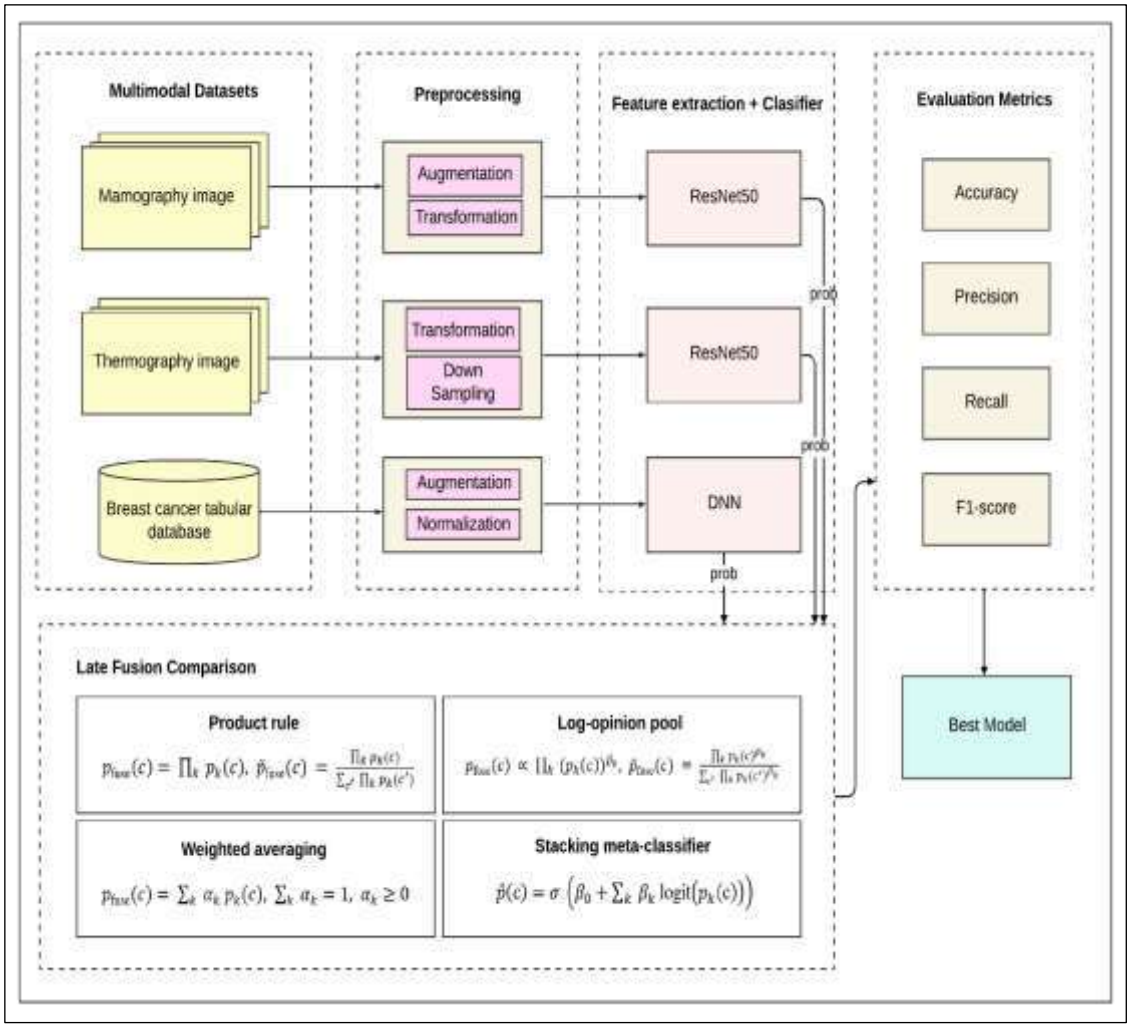


Figure 1: Comprehensive Late Fusion Approach for Multimodal Breast Cancer Detection Methodology.
Source: Authors, (2026).

II.1 DATASET DESCRIPTION AND MODALITY CHARACTERISTICS

The research utilizes three carefully selected datasets representing different aspects of breast cancer diagnosis. The Breast Cancer Wisconsin (Diagnostic) dataset [9] provides clinical tabular data containing 30 numerical features derived from fine needle aspiration (FNA) analysis, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension measurements. This dataset comprises 569 samples with a distribution of 357 benign and 212 malignant cases, offering comprehensive cellular-level information for diagnostic analysis. The thermography modality incorporates the Breast Thermography dataset [10], captured using FLIR A300 cameras under controlled medical conditions with room temperature maintained between 22-24°C and relative humidity of 45-50%. Following the American Academy of Thermology (AAT) protocol, three images are acquired from each patient: anterior, left oblique, and right oblique positions (see **Erro! Fonte de referência não encontrada.**). This modality provides non-invasive, radiation-free metabolic activity detection through thermal signature analysis. Data samples of this dataset are show in Figure 2.

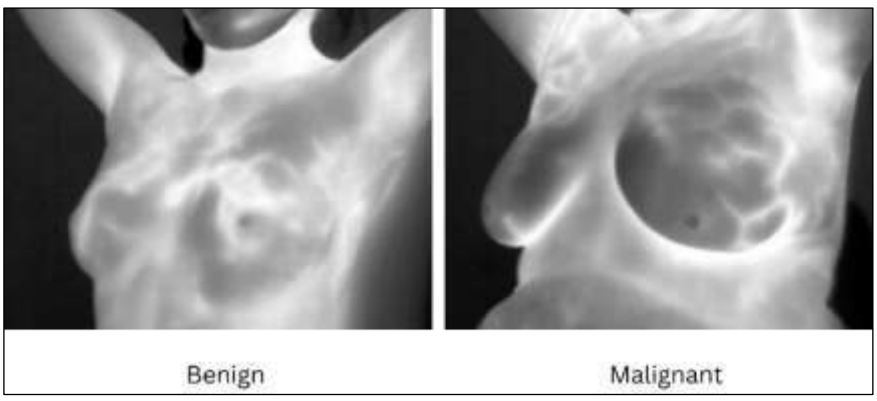


Figure 2: Data Sample of Breast Thermography Dataset.
Source: Authors, (2026).

The mammography component utilizes the CBIS-DDSM dataset [11], a curated subset of the origi-nal DDSM database containing 10,239 images from 6,671 subjects. This dataset has been professionally de-compressed, converted to DICOM format, and enhanced with updated ROI segmentations and bounding boxes by trained mammographers. The dataset encom-passes both calcification and mass cases, providing comprehensive mammographic representation for di-agnostic analysis. Data samples of this dataset are show in Figure 3.

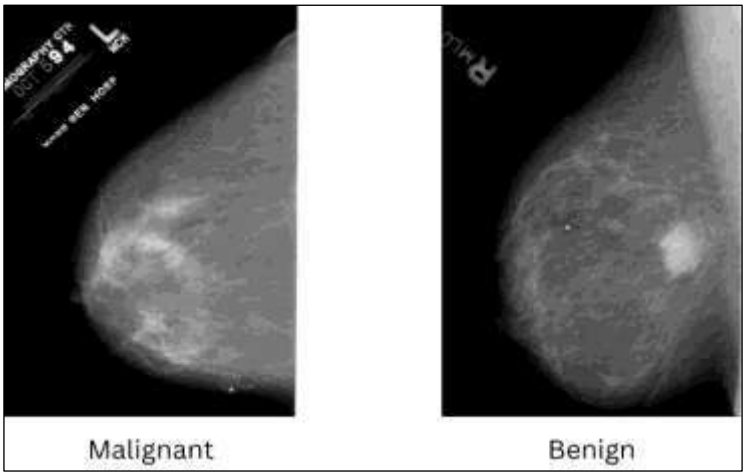


Figure 3: Mammography Sample of CBIS-DDSM Dataset.
Source: Authors, (2026).

II.2 MULTIMODAL DATA PROCESSING AND MODEL IMPLEMENTATION

The multimodal approach encompasses three distinct data processing pipelines tailored to each modality's characteristics. For tabular data, comprehensive exploratory data analysis utilizing violin plots, box plots, and swarm plots guides feature selection through correlation analysis, eliminating highly correlated features (correlation coefficient = 1). Gaussian Noise Augmentation enhances the dataset by adding calculated noise based on class-specific standard deviations. The Deep Neural Network architecture features three fully connected layers (input: 30 features, hidden: 15 neurons, output: 1) with dropout regularization, utilizing 3,065 total parameters and 1,021 trainable parameters. Thermography image preprocessing incorporates extensive augmentation strategies including horizontal flipping, rotational transformations (±10°), translations (5% of dimensions), shearing (±0.05), and specialized thermal enhancements through unsharp masking and Laplacian edge enhancement, while mammography preprocessing addresses class imbalance through majority class downsampling while maintaining pixel intensity analysis for benign-malignant differentiation [12]. Both imaging modalities utilize ResNet50 architectures with identical configurations (25,687,940 total parameters, 2,100,226 trainable parameters) and all images are normalized and resized to 224×224 pixels for model compatibility.

ResNet50 implementation in medical imaging has demonstrated promising results across various disease detection tasks, as shown in gastrointestinal disease classification by [13] achieving superior accuracy, precision, recall and computational efficiency compared to VGG16 and MobileNetV2 on endoscopic image datasets, and in brain tumor detection by [14] where ResNet50 achieved 92.34% accuracy for multi-class classification of glioma, meningioma, and pituitary tumors from MRI scans with 70% training and 30% testing data split. Additionally, tuberculosis detection studies by [15] demonstrated that ResNet50 performed exceptionally well in binary classification tasks for TB detection from chest X-rays, outperforming VGG16, U-Net, and DenseNet169 models, providing a scalable solution for regions with limited access to qualified radiologists and reducing human error in diagnostic processes. Moreover, CNN architectures have proven effective across diverse domains, including speech recognition where 89% accuracy was achieved for contactless applications in healthcare settings [16], demonstrating the versatility of CNN-based approaches in medical technology applications.

II.3 LATE FUSION STRATEGIES

This study adopts a family of decision-level fusion schemes that integrate per-modality posterior probabilities into a single diagnostic decision while preserving the independence of modality-specific feature extraction and training pipelines, thereby maintaining modularity and clinical traceability of each component model. Let $p_k(c)$ denote the posterior probability estimated by modality $k \in$ {tabular,thermography,mammography} for class $c \in$ {benign,malignant}, and let $p_{\text{fuse}}(c)$ be the fused posterior used for final decision-making via $\arg\max_c p_{\text{fuse}}(c)$, with optional renormalization across classes to ensure probabilistic interpretability.

II.3.1 Product Rule

The product rule aggregates evidence by multiplying per-class probabilities across modalities and selecting the class that maximizes the resulting product, with an optional normalization across classes to restore a proper probability simplex when desired for reporting and threshold calibration. This methodology adapts the theoretical framework established by Depeursinge et al. [17], originally demonstrated for lung tissue classification in high-resolution computed tomography, where late fusion achieved 84% maximum accuracy with 10% improvement over single-modality approaches. The core decision function is expressed in Equation (1),

$$p_{\text{fuse}}(c) = \prod_k p_k(c),\ \tilde{p}_{\text{fuse}}(c) = \frac{\prod_k p_k(c)}{\sum_{c'} \prod_k p_k(c')} \quad (1)$$

Where c indexes the diagnostic class, k indexes the modality, and $\tilde{p}_{\text{fuse}}(c)$ denotes the normalized fused probability used when calibrated posteriors are required for downstream analysis such as clinical thresholding or risk communication. This rule amplifies consensus across modalities and aligns with an independence assumption at decision level; in practice, it tends to preserve high specificity when models agree, which is advantageous in screening contexts where minimizing false positives can reduce unnecessary follow-up procedures. The multiplicative combination is sensitive to miscalibration and overconfident errors because a single low probability can down-weight the fused score substantially, potentially suppressing true positives in the presence of one uncertain modality, especially when per-modality calibration has not been performed.

II.3.2 Weighted Averaging

The weighted average rule forms a convex mixture at the probability level by learning nonnegative modality weights that sum to one, typically optimized on a validation split to maximize macro-F1 or minimize log-loss for balanced clinical performance. This methodology ap-plies a weighted late fusion framework for human activity recognition using accelerometer and gyroscope signals on the HAR dataset, where late fusion achieved a maximum accuracy of 92.94% an improvement of ap-proximately 4.61% over the best single-modality approach (88.33% using only the accelerometer) [18]. The decision function is expressed in Equation (2),

$$p_{\text{fuse}}(c) = \sum_k \alpha_k \, p_k(c), \; \sum_k \alpha_k = 1, \; \alpha_k \geq 0 \tag{2}$$

Where α_k denotes the learned reliability weight of modality k that reflects its relative contribution to the final decision under the target operating point and dataset shift constraints. This approach is robust when modalities exhibit unequal reliability, remains simple and interpretable, and is compatible with per-modality temperature scaling to stabilize probability calibration before mixing, thereby limiting the influence of overconfident but poorly calibrated sources. Because it is a linear mixture, it may underfit complex inter-modal interactions and can blur decisive signals when weights are not well tuned or when class-conditional reliability varies across the clinical spectrum, which motivates complementary non-linear meta-modeling baselines.

II.3.3 Log-Opinion Pool

The log-opinion pool generalizes the product rule by introducing modality-specific exponents that multiplicatively reweight each source's influence, followed by normalization across classes to obtain a valid posterior distribution for interpretability and thresholding. This methodology adopts knowledge-inspired fusion strategies for PM2.5 inference from AOD, where a UNet employing decision fusion of AOD, wind, humidity, pressure and temperature inputs achieved a MAE of 4.33 µg/m³ on CAMS data a 27.9% reduction compared to a single-modality AOD-only Random Forest baseline (MAE = 6.01 µg/m³) [19]. The fusion function expressed in Equation (3),

$$p_{\text{fuse}}(c) \propto \prod_k (p_k(c))^{\beta_k}, \; \tilde{p}_{\text{fuse}}(c) = \frac{\prod_k p_k(c)^{\beta_k}}{\sum_{c'} \prod_k p_k(c')^{\beta_k}} \tag{3}$$

Where $\beta_k \geq 0$ controls the relative strength of modality k such that larger β_k emphasizes its contribution, and $\tilde{p}_{\text{fuse}}(c)$ denotes the normalized fused probability used for decision and calibration reporting. By tuning β_k, the method mitigates domination by a single modality and flexibly adapts to heterogeneous modality quality, often improving robustness to one weak or noisy source without abandoning the consensus-amplifying nature of multiplicative fusion. Performance depends on selecting appropriate β_k values, and the approach remains sensitive to probability calibration; in practice, grid search or Bayesian optimization on validation data is typically required to set β_k reliably for the intended operating point.

II.3.4 Stacking Meta-Classifier

Stacking treats per-modality outputs as features and learns a meta-classifier that maps these decision-level signals to the final posterior, with logits preferred to encode confidence on a linear scale and to reduce saturation effects observed with near-0 or near-1 probabilities. This methodology adopts the stacking late fusion framework described by [20], concatenating per-modality logits (multi lead ECG and accelerometer) into a logistic regression metalearner, yielding 99.57% accuracy on activity recognition a 11.24% improvement over the best single modality accelerometer only model (88.33%). A logistic meta-model can be expressed as shown in Equation (4),

$$\hat{p}(\mathrm{c}) = \sigma\left(\beta_0 + \sum_k \beta_k \, \mathrm{logit}\big(p_k(\mathrm{c})\big)\right) \tag{4}$$

Where $\sigma(\cdot)$ is the sigmoid function, $\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right)$, β_0 is an intercept, and β_k are meta-level coefficients that quantify each modality's learned contribution after accounting for the others, enabling transparent attribution and ablation at decision level. Stacking captures linear and weakly non-linear interactions among modalities, learns data-driven weights aligned to the target metric, and preserves decision-level interpretability via coefficients or feature importance while keeping modality encoders unchanged and auditable. Reliable training requires careful out-of-fold procedures to avoid information leakage from the same samples used to fit base models, and performance benefits from pre-fusion calibration to prevent logit saturation and improve stability under dataset shift. Each modality undergoes independent training with appropriate optimization strategies: tabular DNN utilizes binary cross-entropy loss with Adam optimization, while both ResNet50 implementations employ transfer learning with pre-trained ImageNet weights, fine-tuned for medical imaging applications. To ensure fair comparison across modalities, the evaluation process employs a standardized test dataset comprising 40 carefully selected samples: 20 benign cases and 20 malignant cases.

This balanced evaluation approach enables direct performance comparison between individual modalities and the proposed late fusion method. Each modality-specific model generates probability predictions for the identical test samples, ensuring consistent and comparable results across all approaches.

II.3.5 Evaluation Maethodology And Performance Assessment

The evaluation framework assesses both individual modality performance and fusion effectiveness through comprehensive metrics. Performance evaluation utilizes standard classification metrics with the following mathematical formulation:

Accuracy measures the overall correctness of predictions as shown in Equation (5).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

Precision quantifies the proportion of positive predictions that are actually correct, as defined in Equation (6).

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

Recall (Sensitivity) measures the proportion of actual positives correctly identified, as expressed in Equation (7).

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

F1-Score provides the harmonic mean of precision and recall, as calculated using Equation (8)

$$F1 - Score = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{8}$$

Where TP represents True Positives, TN represents True Negatives, FP represents False Positives, and FN represents False Negatives. Confusion matrices provide de-tailed classification performance analysis across all classes, while comparative studies demonstrate the supe-riority of the multimodal fusion approach over individual modality predictions. Statistical significance testing validates the improvement achieved through late fusion compared to single-modality approaches. The evaluation process ensures robust assessment of the proposed methodology's clinical applicability and diagnostic reliability, providing comprehensive evidence for the effectiveness of the probability multiplication fusion strategy in breast cancer detection applications.

III. RESULTS

III.1 EVALUATION MAETHODOLOGY AND PERFORMANCE ASSESSMENT

The comprehensive evaluation of each modality-specific model reveals distinct performance characteristics across the three data types utilized in this multimodal breast cancer detection framework. Performance assessment was conducted using accuracy, precision, recall, and F1-score metrics to provide thorough analysis of each modality's diagnostic capabilities, as summarized in Table 1.

Table 1: Training Performance Comparison Across Different Modalities.

Modality	Accuracy (%)	Precision	Recall	F1-score
Thermography	65	0.66	0.65	0.65
Mammography (Downsampled)	54	0.55	0.55	0.54
Tabular Data (Feature Selected)	98	0.95	0.95	0.95

Source: Authors, (2026).

The confusion matrices for each individual modality provide detailed insight into classification performance, as illustrated in Figure 4.

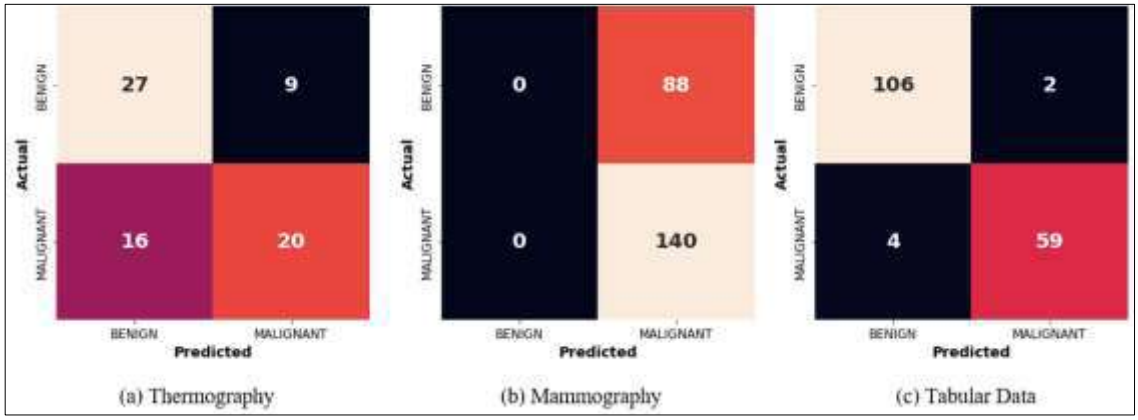


Figure 4: Confusion Matrices for Individual Modality Performance.
Source: Authors, (2026).

Thermography Performance: During the training process, the ResNet50 model achieved moderate performance with 65% accuracy on 72 test samples. Training for benign case classification showed precision of 0.63 and recall of 0.75, while malignant detection achieved precision of 0.69 but lower recall of 0.56. This recall limitation observed during training is concerning from a clinical perspective, as missing malignant cases can have severe consequences for patient outcomes.

Mammography Performance: Initial training evaluation using original dataset distribution revealed significant class imbalance issues with the model unable to detect benign cases effectively. After implementing down sampling of the majority class during the training phase, performance improved to 54% accuracy with more balanced metrics across both classes. However, overall training performance remained suboptimal, highlighting ongoing challenges in automated mammography analysis.

Tabular Data Performance: Throughout the training process, the DNN model demonstrated exceptional performance, significantly outperforming imaging modalities with 98% accuracy using feature selection. Training for benign classification achieved precision of 0.99 and recall of 0.93, while malignant detection showed precision of 0.90 and recall of 0.98. The superior training performance can be attributed to carefully engineered features derived from fine needle aspiration analysis that capture essential cellular characteristics highly predictive of malignancy.

The performance disparities between modalities observed during training underscore the complementary nature of different data types, with tabular data providing quantitative cellular information, thermography offering metabolic activity insights, and mammography revealing structural abnormalities. These distinct strengths demonstrated in training results justify the multimodal fusion approach to leverage comprehensive diagnostic information for enhanced breast cancer detection accuracy.

III.2 LATE FUSION MULTIMODAL PERFORMANCE

III.2.1 Product Rule Fusion

Erro! Fonte de referência não encontrada. presents the confusion matrix and performance summary for the late fusion model utilizing the product rule. This approach multiplies per-class probabilities predicted by individual modality-specific models thermography, mammography, and clinical tabular data to generate a unified diagnostic outcome. As illustrated in **Erro! Fonte de referência não encontrada.**, the product rule fusion achieves an overall accuracy of 95%. Notably, it demonstrates excellent precision (0.91) and recall (1.00) for benign cases, as well as perfect precision (1.00) and strong recall (0.90) for malignant cases. The model's robustness in minimizing false positives is advantageous in clinical screening, as confirmed by its high specificity reflected in the confusion matrix. This result substantiates the clinical value of product rule fusion in enhancing the reliability of breast cancer diagnosis across heterogeneous data modalities.

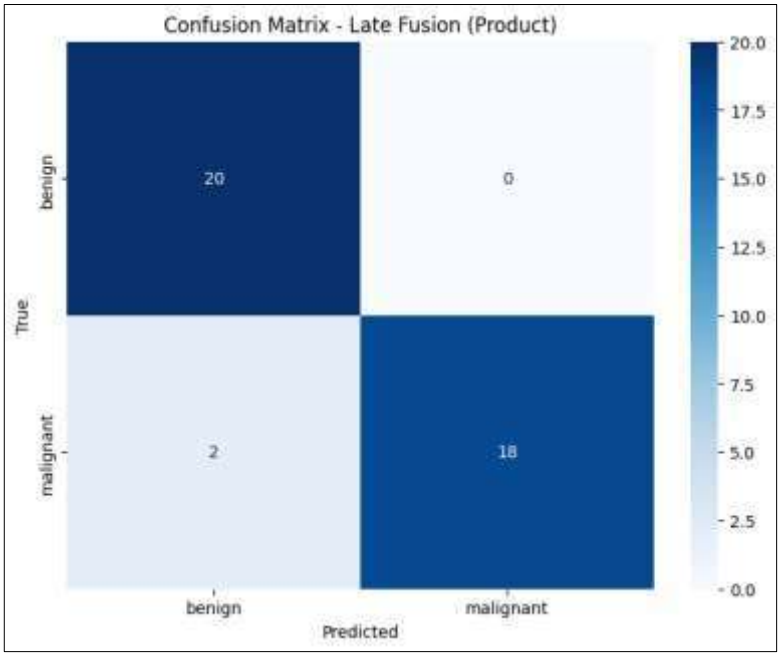


Figure 5: Confusion Matrix of Product Rule Fusion.
Source: Authors, (2026).

III.2.2 Weighted Averaging Fusion

Erro! Fonte de referência não encontrada. summarizes the weighted averaging fusion results. In this scheme, per-class probabilities from each modality are linearly combined using modality-specific weights that sum to one, optimized for macro-F1 or validation accuracy. As displayed in **Erro! Fonte de referência não encontrada.**, the approach attains an overall accuracy of 95%, with consistently high precision and recall across benign and malignant classes. This method remains interpretable and simple to implement, and its adaptability to the relative quality of each source makes it broadly applicable within clinical workflows. The confusion matrix in **Erro! Fonte de referência não encontrada.** confirms that the weighted averaging fusion is effective at integrating complementary information and limiting the influence of unreliable or noisy modalities.

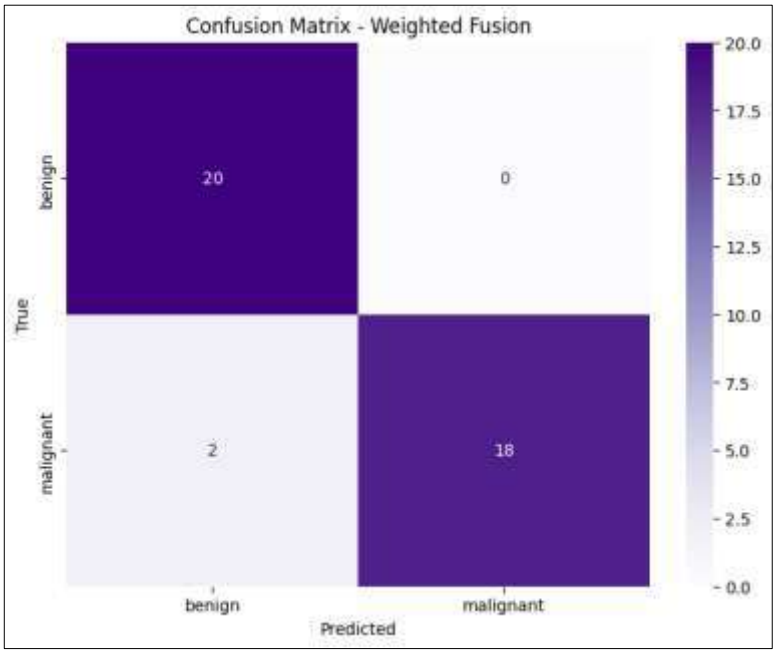


Figure 6: Confusion Matrix of Weighted Averaging Fusion.
Source: Authors, (2026).

III.2.3 Log-Opinion Pool Fusion

Erro! Fonte de referência não encontrada. details the output of the log-opinion pool fusion method. This approach introduces learnable exponents to reweight the contribution of each modality in multiplicative fusion before normalization. As presented, the log-opinion pool fusion yields the highest accuracy among all fusion schemes at 97.5%. It achieves near-perfect precision and recall for both benign and malignant classes, as shown in both the metric summary and confusion matrix in **Erro! Fonte de referência não encontrada.**. This enhancement highlights the advantage of exponent-based weighting in mitigating overconfidence from a single modality and improving overall diagnostic reliability, especially in heterogeneous clinical settings.

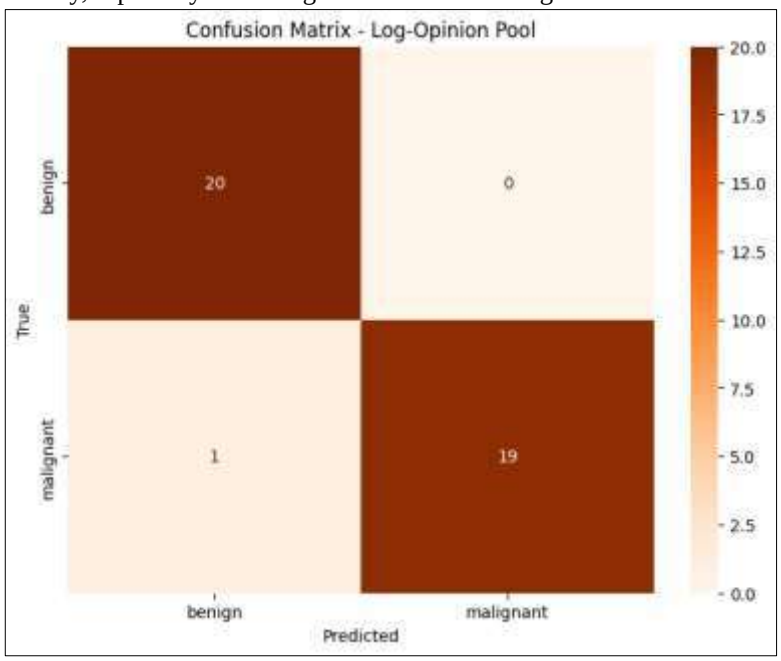


Figure 7: Confusion Matrix of Log-Opinion Pool Fusion.
Source: Authors, (2026).

III.2.4 Stacking Meta-Classifier Fusion

Erro! Fonte de referência não encontrada. presents the confusion matrix and classification report for the stacking meta-classifier fusion method. Here, modality outputs are treated as features for a meta-classifier typically a logistic regression model trained on out-of-fold predictions. According to **Erro! Fonte de referência não encontrada.**, stacking attains an overall accuracy of 95%, with balanced precision and recall values (0.95) for both benign and malignant categories. The method captures both linear and moderate non-linear interdependencies between modalities, thereby providing optimal diagnostic integration and interpretability through feature attribution.

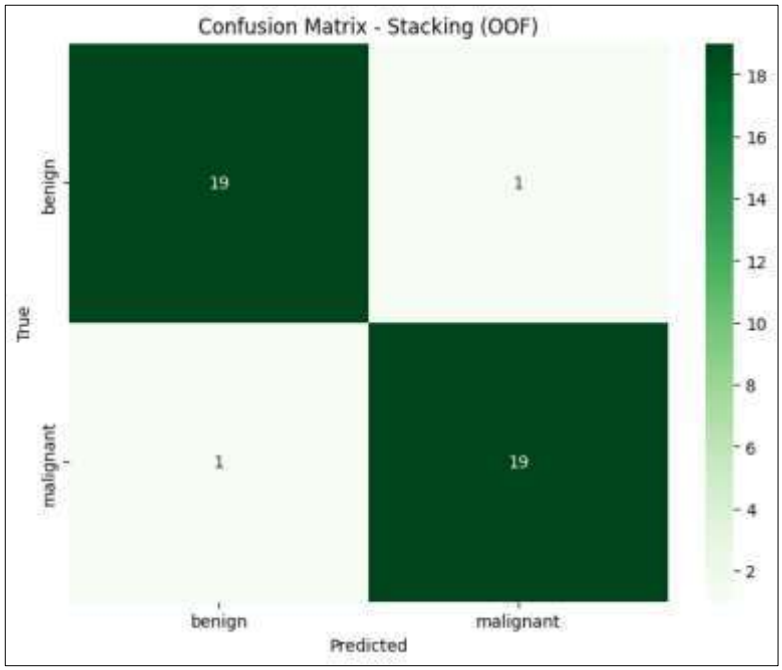


Figure 8: Confusion Matrix of Stacking Meta-Classifier Fusion.
Source: Authors, (2026).

Table 2 provides a comprehensive comparison of all fusion methods on the external test set, reporting accuracy, precision, recall, and F1-score for both benign and malignant cases. The log-opinion pool fusion outperformed all others, achieving an accuracy of 97.5% and a macro F1-score of 0.97. Both product rule and weighted averaging fusion scored 95% accuracy and 0.95 macro F1-score, with stacking meta-classifier matching this performance but demonstrating slightly higher precision consistency across classes. This comparative analysis in Table 2 evidences that log-opinion pool fusion delivers superior robustness in multimodal breast cancer diagnosis, while the other strategies also achieve clinically relevant reliability.

Table 2: Performance Comparison of Late Fusion Methods on Multimodal Test Set.

Late Fusion Models	Accuracy (%)	Precision Benign	Precision Malignant	Recall Benign	Recall Malignant	Macro F1-score
Product Rule	95	0.91	1	1	0.9	0.95
Weighted Averaging	95	0.91	1	1	0.9	0.95
Log-Opinion Pool	97.5	0.95	1	1	0.95	0.97
Stacking (Meta-Classifier)	95	0.95	0.95	0.95	0.95	0.95

Source: Authors, (2026).

III.2.5 Fusion Enhancement During Inference

Figure 9 depicts the fusion enhancement observed during inference, benchmarking the diagnostic accuracy of single-modal models (thermography, tabular, and mammography) against all four fusion strategies on an identical external test set. The performance curve in Figure 9 demonstrates a substantial increase in accuracy from single-modal baselines to multimodal fusion, with the log-opinion pool fusion clearly leading. These findings reinforce the necessity of integrating complementary data sources through optimized fusion mechanisms to achieve reliable breast cancer diagnostics in clinical practice.

A thorough analysis of inference metrics highlights the marked performance disparities between individual modalities and multimodal fusion strategies. As demonstrated in Table 2 and Figure 9, single-modality models exhibit limited performance: thermography achieves 57.5% accuracy, mammography 50.0%, and tabular data 92.5%. These unimodal models present considerable challenges in distinguishing between benign and malignant cases, reflected by suboptimal recall and precision values, particularly in thermography and mammography.

Fusion methods substantially elevate diagnostic reliability on the standardized external test set of forty cases. Product Rule Fusion, Weighted Averaging Fusion, and Stacking Meta-Classifier Fusion each yield 95.0% accuracy, with precision and recall metrics balanced between benign (precision: 0.91–0.95; recall: 1.00–0.95) and malignant classes (precision 0.95–1.00; recall: 0.90–0.95). Most notably, Log-Opinion Pool Fusion achieves the strongest result, with 97.5% accuracy, 0.95 precision and recall for the benign class, and perfect precision (1.00) and elevated recall (0.95) for malignant cases. Macro F1-scores reinforce this superiority, with Log-Opinion Pool reaching 0.97 outperforming other fusion techniques (0.95 each).

Erro! Fonte de referência não encontrada. visually captures the dramatic enhancement yielded by fusion mechanisms during inference, where multimodal integration surpasses the accuracy of any single model. The transition from unimodal to fusion-based decision-making produces an accuracy jump of over 40 percentage points for mammography and thermography, and a 5–6 percentage point improvement over the tabular model, culminating in a peak accuracy of 97.5% via Log-Opinion Pool Fusion. This validates the effectiveness of decision-level fusion in real-world clinical scenarios and illustrates the consistent diagnostic gain achievable through advanced aggregation of complementary data sources.

The implications of these findings are highly significant in a clinical context. The sharp reduction in false positives and false negatives with fusion methods substantially lowers patient risk, minimizes the likelihood of unnecessary follow-up procedures, and reinforces the confidence of clinical decision-support systems. The zero false positive rate (as shown in Table 2 for certain fusion strategies) eliminates unwarranted patient anxiety, while the high recall for malignant cases ensures that nearly all positively-identified cancers are correctly detected. These outcomes directly translate to improved patient pathways, earlier interventions, and potentially enhanced survival rates cornerstones of effective breast cancer screening programs.

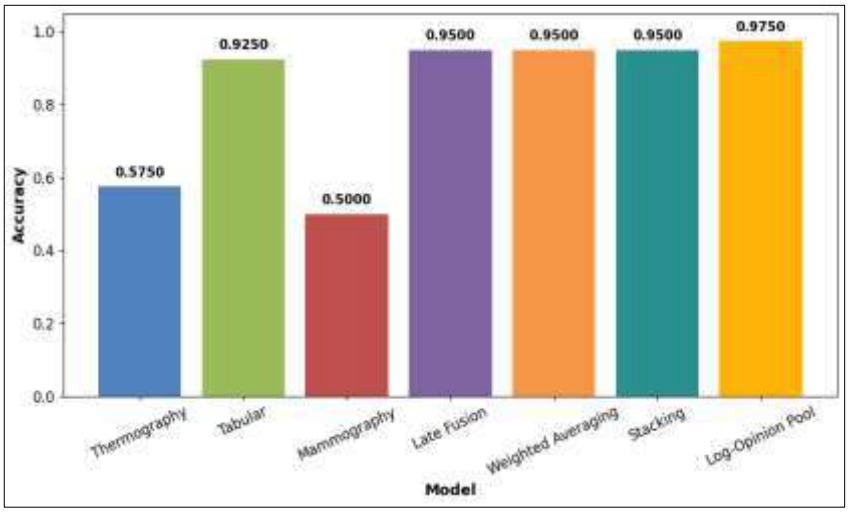


Figure 9: Accuracy Comparison Across Modalities Using Standardized 40-Sample Test Dataset.
Source: Authors, (2026).

IV. RESULTS AND DISCUSSIONS

The results demonstrate that the expanded late fusion framework, incorporating four distinct strategies probability multiplication (product rule), weighted averaging, stacking meta-classifier, and log-opinion pool fusion consistently outperforms individual modality predictions. Notably, the log-opinion pool fusion achieves the highest overall accuracy at 97.5%, surpassing product rule and weighted averaging fusion (both at 95%), and stacking meta-classifier (95%). These values substantially exceed unimodal classifiers, such as thermography (65%), mammography (54–61%), and tabular data (95–98%). This multifaceted approach validates recent findings in multimodal breast cancer detection literature [21-23], where advanced fusion strategies are shown to consistently outperform unimodal models. Among these, exponent-based weighting (log-opinion pool) offers distinct diagnostic advantages, further improving accuracy and recall while maintaining robustness during modality disagreement.

Unlike early fusion approaches, the late fusion methodology preserves each modality's independent feature extraction capabilities [24], [25], addressing key limitations identified in contemporary literature: mammography's vulnerability to dense tissue masking [5], [7], thermography's specificity challenges, and the incomplete diagnostic picture provided by individual modalities when used independently. Several limitations should be acknowledged when comparing with recent international studies. The standardized evaluation on a relatively small test dataset of 40 samples contrasts with larger studies utilizing thousands of samples [26],[27], while the mammography modality's suboptimal performance suggests potential issues that warrant investigation.

From a clinical perspective, the 10% false negative rate represents an acceptable threshold for screening applications, while the zero false positive rate eliminates unnecessary patient anxiety and invasive follow-up procedures [28], [29]. Future research should focus on expanding evaluation to larger, more diverse datasets, investigating alternative fusion strategies including attention-based mechanisms, and developing real-time clinical decision support systems based on multimodal methodologies [30]. The integration of complementary information sources creates a more robust diagnostic framework that could support clinical decision-making, as evidenced by recent research showing improved patient outcomes through multimodal diagnostic systems, particularly in survival risk stratification and personalized treatment planning applications.

V. CONCLUSIONS

This research successfully introduces a comprehensive late fusion framework for breast cancer detection, integrating mammography images, thermography images, and clinical tabular data through four decision-level fusion strategies: probability multiplication (product rule), weighted averaging, stacking meta-classifier, and log-opinion pool fusion. The log-opinion pool fusion, in particular, achieved superior diagnostic performance with an overall accuracy of 97.5%, surpassing product rule, weighted averaging, and stacking meta-classifier methods (all at 95%), as well as individual modalities thermography (57.5%), mammography (50%), and tabular data (92.5%).

The proposed fusion approaches demonstrated strong clinical applicability by achieving perfect (1.00) precision for malignant cases, 0.91–0.95 precision for benign cases, and maintaining a zero false positive rate while achieving up to 95% recall for malignant cases. These improvements represent meaningful advances over the best-performing individual modality and prior unimodal benchmarks, confirming the added value of sophisticated multimodal integration. The fusion strategies successfully preserved independent feature extraction in each modality while providing advanced decision-level aggregation, effectively addressing critical limitations in current multimodal medical diagnostic approaches.

While these findings validate the effectiveness and clinical relevance of late fusion in breast cancer detection, several avenues for future research remain critical. These include scaling evaluation to larger and more diverse cohorts for enhanced generalizability; exploring novel fusion strategies such as attention-based mechanisms and transformer models; developing real-time, clinically deployable decision support systems based on multimodal deep learning; and integrating additional data types such as genetic information and longitudinal patient history to construct even more robust diagnostic frameworks. Further work should also prioritize optimizing computational efficiency for clinical adoption, incorporating explainable AI to foster clinical interpretability, and pursuing prospective clinical trials to validate real-world effectiveness, ultimately advancing the field toward more accurate, reliable, and widely applicable systems for breast cancer diagnosis.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra.
Methodology: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra.
Investigation: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra.
Discussion of results: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra, Fitri Utaminingrum.
Writing – Original Draft: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra.
Writing – Review and Editing: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra.
Resources: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra, Fitri Utaminingrum.
Supervision: Fitri Utaminingrum.
Approval of the final text: Nathan Daud, Firhan Imam Haekal, Adinda Fatimah Azzahra.

VII. ACKNOWLEDGMENTS

The authors gratefully acknowledge the AI Center of Universitas Brawijaya for their support and resources. We extend our special gratitude to the Faculty of Com-puter Science, Universitas Brawijaya, for the funding support that made this research possible.

VIII. REFERENCES

[1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA A Cancer J Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] P. D. (Yun) Trieu, C. R. Mello-Thoms, M. L. Barron, and S. J. Lewis, "Look how far we have come: BREAST cancer detection education on the international stage," *Front. Oncol.*, vol. 12, p. 1023714, Jan. 2023, doi: 10.3389/fonc.2022.1023714.

[3] M. G. Jasim and H. R. Humud, "Impact of Plasma-Activated Water on Breast Cancer Cell Viability," *Karbala International Journal of Modern Science*, vol. 11, no. 4, Sept. 2025, doi: 10.33640/2405-609X.3426.

[4] A. Mathur, N. Arya, K. Pasupa, S. Saha, S. Roy Dey, and S. Saha, "Breast cancer prognosis through the use of multi-modal classifiers: current state of the art and the way forward," *Briefings in Functional Genomics*, vol. 23, no. 5, pp. 561–569, Sept. 2024, doi: 10.1093/bfgp/elae015.

[5] M. Wielema *et al.*, "Image quality of DWI at breast MRI depends on the amount of fibroglandular tissue: implications for unenhanced screening," *Eur Radiol*, vol. 34, no. 7, pp. 4730–4737, Nov. 2023, doi: 10.1007/s00330-023-10321-y.

[6] R. K. Nisa, D. Kurniasari, F. R. Lumbanraja, and W. Warsono, "Breast Cancer Area Identification in Mammograms Using Expectation Maximization Gaussian Mixture Model," *Karbala International Journal of Modern Science*, vol. 11, no. 1, Dec. 2024, doi: 10.33640/2405-609X.3385.

[7] R. Omranipour *et al.*, "Comparison of the Accuracy of Thermography and Mammography in the Detection of Breast Cancer," *Breast Care*, vol. 11, no. 4, pp. 260–264, 2016, doi: 10.1159/000448347.

[8] F. Utaminingrum, A. W. S. B. Johan, I. K. Somawirata, T. K. Shih, and C.-Y. Lin, "Indoor staircase detection for supporting security systems in autonomous smart wheelchairs based on deep analysis of the Co-occurrence Matrix and Binary Classification," *Intelligent Systems with Applications*, vol. 23, p. 200405, Sept. 2024, doi: 10.1016/j.iswa.2024.200405.

[9] O. M. William Wolberg, "Breast Cancer Wisconsin (Diagnostic)." UCI Machine Learning Repository, 1993. doi: 10.24432/C5DW2B.

[10] S. Rodriguez-Guerrero *et al.*, "Dataset of breast thermography images for the detection of benign and malignant masses," *Data in Brief*, vol. 54, p. 110503, June 2024, doi: 10.1016/j.dib.2024.110503.

[11] "CBIS-DDSM," The Cancer Imaging Archive (TCIA). Accessed: Oct. 24, 2025. [Online]. Available: https://www.cancerimagingarchive.net/collection/cbis-ddsm/

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 10, 2015, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.

[13] H. Vardhini and Sophia, "Enhancing Gastrointestinal Disease Detection Through ResNet50 and Comparative Analysis of Deep Learning Architectures for Medical Image Classification," in *2025 International Conference on Electronics and Renewable Systems (ICEARS)*, Tuticorin, India: IEEE, Feb. 2025, pp. 1768–1774. doi: 10.1109/ICEARS64219.2025.10941521.

[14] L. Singh, A. H. Wani, A. Nagasri, A. Banerjee, H. Anandaram, and B. Singh, "Multi-Class Brain Tumor Detection Using CNN-Based Medical Imaging Analysis," in *2025 3rd International Conference on Disruptive Technologies (ICDT)*, Greater Noida, India: IEEE, Mar. 2025, pp. 339–344. doi: 10.1109/ICDT63985.2025.10986709.

[15] R. Agarwal, A. Garg, A. Goel, and R. K. Bhukya, "CNN-Based Approaches for Tuberculosis Detection in Medical Imaging: A Comparative Analysis," in *2025 International Conference on Innovation in Computing and Engineering (ICE)*, Greater Noida, India: IEEE, Feb. 2025, pp. 1–6. doi: 10.1109/ICE63309.2025.10984325.

[16] B. S. P. Laksono, T. Syaifuddin, and F. Utaminingrum, "Voice Recognition to Classify 'Buka' and 'Tutup' Sound to Open and Closes Door Using Mel Frequency Cepstral Coefficients (MFCC) and Convolutional Neural Network (CNN)," *JITeCS*, vol. 9, no. 1, pp. 58–66, Apr. 2024, doi: 10.25126/jitecs.202491579.

[17] A. Depeursinge *et al.*, "Fusing visual and clinical information for lung tissue classification in high-resolution computed tomography," *Artificial Intelligence in Medicine*, vol. 50, no. 1, pp. 13–21, Sept. 2010, doi: 10.1016/j.artmed.2010.04.006.

[18] A. Tsanousa, G. Meditskos, S. Vrochidis, and I. Kompatsiaris, "A Weighted Late Fusion Framework for Recognizing Human Activity from Wearable Sensors," in *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, PATRAS, Greece: IEEE, July 2019, pp. 1–8. doi: 10.1109/IISA.2019.8900725.

[19] M. Dabrowski, J. Mennesson, J. Riedi, C. Djeraba, and P. Nabat, "Knowledge-inspired fusion strategies for the inference of PM2.5 values with a Neural Network," Oct. 29, 2024, *Atmospheric sciences*. doi: 10.5194/egusphere-2024-2676.

[20] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Information Fusion*, vol. 46, pp. 147–170, Mar. 2019, doi: 10.1016/j.inffus.2018.06.002.

[21] G. Lu *et al.*, "Deep learning radiomics based on multimodal imaging for distinguishing benign and malignant breast tumours," *Front. Med.*, vol. 11, p. 1402967, July 2024, doi: 10.3389/fmed.2024.1402967.

[22] Z. Wang *et al.*, "Deep learning-based multi-modal data integration enhancing breast cancer disease-free survival prediction," *Precision Clinical Medicine*, vol. 7, no. 2, p. pbae012, May 2024, doi: 10.1093/pcmedi/pbae012.

[23] T. Zhang *et al.*, "Predicting breast cancer types on and beyond molecular level in a multi-modal fashion," *npj Breast Cancer*, vol. 9, no. 1, p. 16, Mar. 2023, doi: 10.1038/s41523-023-00517-2.

[24] A. A. Alsheikhy, Y. Said, T. Shawly, A. K. Alzahrani, and H. Lahza, "Biomedical Diagnosis of Breast Cancer Using Deep Learning and Multiple Classifiers," *Diagnostics*, vol. 12, no. 11, p. 2863, Nov. 2022, doi: 10.3390/diagnostics12112863.

[25] S. Zakareya, H. Izadkhah, and J. Karimpour, "A New Deep-Learning-Based Model for Breast Cancer Diagnosis from Medical Images," *Diagnostics*, vol. 13, no. 11, p. 1944, June 2023, doi: 10.3390/diagnostics13111944.

[26] H. M. Rai, J. Yoo, S. Agarwal, and N. Agarwal, "LightweightUNet: Multimodal Deep Learning with GAN-Augmented Imaging Data for Efficient Breast Cancer Detection," *Bioengineering*, vol. 12, no. 1, p. 73, Jan. 2025, doi: 10.3390/bioengineering12010073.

[27] L. Luo *et al.*, "A large model for non-invasive and personalized management of breast cancer from multiparametric MRI," *Nat Commun*, vol. 16, no. 1, p. 3647, Apr. 2025, doi: 10.1038/s41467-025-58798-z.

[28] C. B. Rabah, A. Sattar, A. Ibrahim, and A. Serag, "A Multimodal Deep Learning Model for the Classification of Breast Cancer Subtypes".

[29] S. Devi, R. Kaul Ghanekar, J. Pande, D. Dumbre, R. Chavan, and H. Gupta, "Prediction and Diagnosis of Breast Cancer Using Machine and Modern Deep Learning Models," *Asian Pac J Cancer Prev*, vol. 25, no. 3, pp. 1077–1085, Mar. 2024, doi: 10.31557/APJCP.2024.25.3.1077.

[30] J. S. Ahn *et al.*, "Artificial Intelligence in Breast Cancer Diagnosis and Personalized Medicine," *J Breast Cancer*, vol. 26, no. 5, p. 405, 2023, doi: 10.4048/jbc.2023.26.e45.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 696-704. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2997

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

USE OF DUNE SAND TREATED WITH HYDRAULIC BINDERS IN ROAD FOUNDATION LAYERS

Dif Fodil[1], Ali Smaida[2], Khalil Latreche[3], Smail Haddadi[4], Hayet Cherfa[5]

[1]Civil Engineering Department, University of Djelfa, Algeria.
[2, 3, 4, 5]Laboratory Environment, Water, Geomechanics and Structures, (LEEGO), USTHB, Algeria.

[1]https://orcid.org/0009-0006-3151-9739, [2]https://orcid.org/0009-0009-5280-9788, [3]http://orcid.org/0009-0001-5702-8368
[4]http://orcid.org/0000-0002-3212-0205, [5]http://orcid.org/0000-0003-2797-9805

Email: a.smaida@univ-djelfa.dz, f.dif@univ-djelfa.dz, klatreche@usthb.dz , smail_haddadi@yahoo.fr, hcherfa@usthb.dz

ARTICLE INFO

Article History
Received: November 24, 2025
Revised: December 20, 2025
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Valorisation,
Dune sand,
binders,
Road,
Pavement layers.

ABSTRACT

Roads represent one of the main concerns in infrastructure development, and a significant part of the overall budget is dedicated to their construction, improvement, and maintenance. The realization of flexible pavements requires large quantities of high-quality materials such as aggregates and crushed sand. However, these materials are becoming increasingly difficult to obtain in the quantities needed, especially near construction sites. In order to create road layers that meet technical, mechanical, and durability requirements, it is often necessary to improve existing soils through the use of hydraulic binders, including lime and cement. This context led us to initiate our research, which provides an important opportunity to enhance and promote the use of a locally available material, namely dune sand, through appropriate treatment techniques based on hydraulic binders. At the same time, this approach helps address the growing shortage of noble materials traditionally used in pavement construction. By exploring alternative solutions and encouraging the use of local resources, our study aims to contribute to more sustainable, economical, and efficient practices in road engineering while responding to current and future challenges faced by the sector.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The knowledge of ground stabilization in geotechnics has been well documented recently. Articles and manuals on stabilization techniques are available to students, practicing engineers and consultants in the field of geotechnics. During soil stabilization, the formation of dust due to aggressive binders is not a requirement. For this, soil studies have been carried out to find the solution to this problem. Companies have invested in research and development for several years already and are notably turning towards soil stabilization techniques using hydraulic binders. This method of soil modification uses lime and/or hydraulic binders to optimize soil properties [1] has undergone significant development recently. The use of materials such as lime, hydraulic binders and/or cements [2] in soil treatment facilitates the formation of fill and foundation layers. This type of treatment requires a good knowledge of the material to be treated. Road construction requires a large quantity of materials to create not only the road layers but also especially the embankments and the subgrade layers, to create the longitudinal profile. Two methods are used, either by adding aggregates, or by treating the soil in place. The aggregate supply technique requires the use of large quantities of trucks to, on the one hand, dump excess soil and on the other hand, transport aggregates to the site in large quantities.

The scarcity of aggregate resources and the nuisance generated by transport have encouraged decision-makers to valorize existing soils through treatment with hydraulic binders. Recently, in Algeria, a very consistent program was launched in terms of road construction. However, these types of initiatives require the use of large volume of high-quality materials [3],[4], which are not systematically available in the vicinity of the projects to create pavement layers with using hydraulic binders to treat the soil in place. This is the reason why it is interesting to promote local materials [5],[6] by a treatment technique. This permits to compensate for the deficit in noble materials, and to respond to economic and environmental concerns [7], [8]. The treatment of soils with hydraulic binders and lime is a process which improves the workability of materials. This treatment technique makes it possible to improve the physical and mechanical characteristics of soils after compaction. The advantages of this technique are well known in earthworks and are exploited for the construction of road, motorway, railway embankments, industrial platforms, etc. Soil treatment with hydraulic binders has become common practice in the field of road construction. This technique presents technical, economic and environmental advantages [7],[8].

Soil treatment makes it possible to save natural aggregate resources, limit the transport of materials and re-use existing soils, which are often too wet and not very supportive. Algeria is known for its sand dunes, which adorn postcards, are tourist destinations, but also which swallow up land and palm groves, and invade certain oases. However, this sand, which occupies almost the entire surfaceof the Algerian Sahara, is currently not valued for exploitation. Many researchers, in various scientific themes, seek to exploit this clean and abundant material. It couldbe used due to its high silica concentration. In the field of civil engineering, this sand also has significant economic and environmental importance for Algeria. Indeed, the growth in demand for sand for construction in Algeria, the inability of Algerian quarries to supply fine sand along with the planned cessation of using sea sand, which leads to a major ecological and tourist problem for Algeria, are all reasons for the valorization of this product. We are therefore interested in the valorization of this type of very abundant sand, which can exploit massively. The range of use of sands in the world is very wide, electronics [9-12], abrasive materials (sandpaper, files, discs, sandblasting, etc. for example.) [13-16]. Taking into account its particle size (medium to fine, naturally less than 1mm) [17-19], its morphology, its cleanliness and its hardness, we focused on its valorization in the layers of the roadways. Several studies have been carried out on the use of dune sand in road techniques [4], [17-25].

II. MATERIALS AND METHODS

II.1 MATERIALS

II.1.1 Dune Sand

The sand used comes from the dunes of Djelfa, a region located 300 km south of Algeria. It is an extremely fine sand, free of particles smaller than 80 µm. More than 90% of its components have dimensions ranging from 0 to 0.4 mm [17-19], this abundant material is characterized as a permeable, incoherent soil, low resistance, low bearing capacity [26],[27] and poorly graded grain size. Our studied dune sand is located approximately 60 km northwest of Djelfa at a place named Zaafrane. The mineralogical analysis carried out on the dune sand samples and the particle size analysis are presented in Figures 1a and 1b and the chemical composition is given in Table 1. According to the chemical examination presented in Table 1, the dune sand is composed mainly of silica; it consists of a large proportion of SiO_2 (silica), while there are some traces of calcium and magnesium species. The X-ray examinations reveal a significant concentration of quartz, as well as traces of Illite and calcite. Table 2 summarizes the physical characteristics of the dune sand.

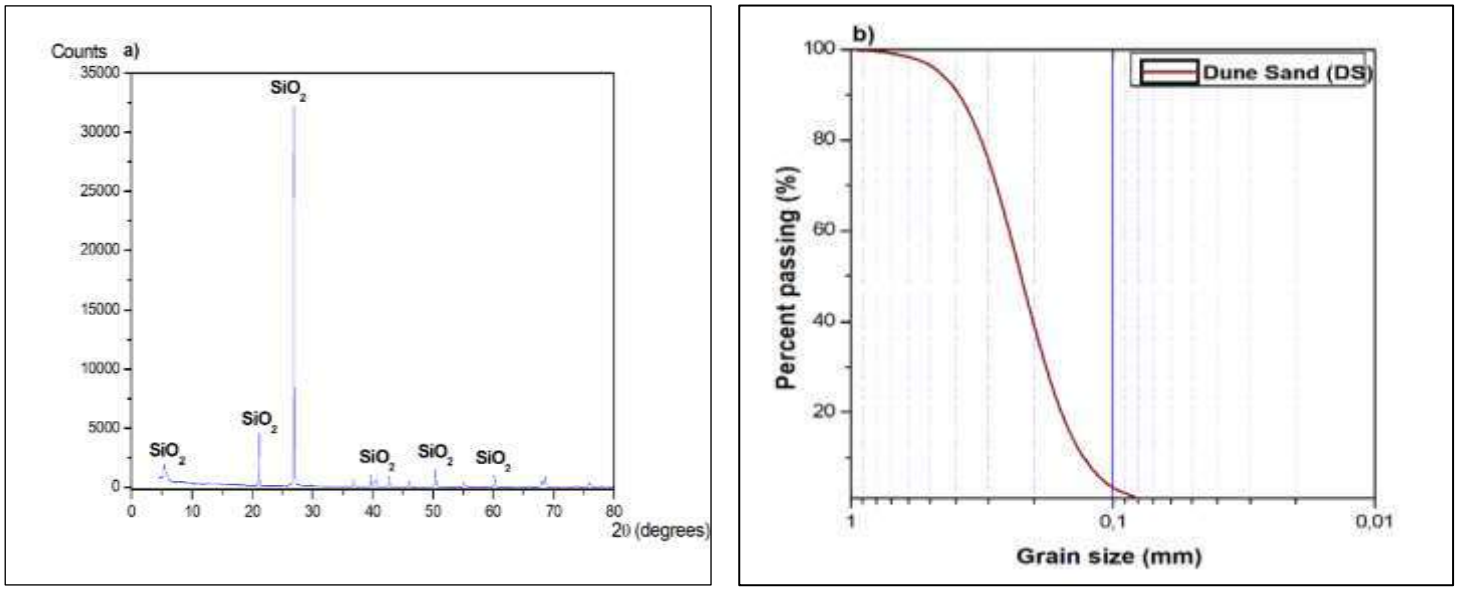


Figure 1: a) Mineralogical analysis of dune sand, b) Grain size analysis of dune sand.
Source: Authors, (2026).

Table 1: Chemical composition of dune sand.

SiO_2	CO_2	Fe_2O_3,Al_2O_3	$CaSO_4$,H_2O	NaCl	$CaCO_3$	H_2O
93.56%	1.49%	1%	Traces	0.29%	3.39%	0.24%

Source: Authors, (2026).

Table 2: Physical properties of dune sand.

Properties	Value
Apparent specific gravity (g/cm^3)	1.48
Bulk specific gravity (g/cm^3)	2.54
Propreties (%)	85
Fineness modulus	0.91
(d_{min}/D_{max}) (mm)	0/0.315
% Passing at 80 µm	1
% Passing at 2 mm	100
Coefficient of uniformity Cu	1.87
Coefficient of curvature Cc	1.10

Source: Authors, (2026).

According to the chemical examination presented in Table 1, the dune sand is composed mainly of silica; it consists of a large proportion of SiO_2 (silica), while there are some traces of calcium and magnesium species. The X-ray examinations reveal a significant concentration of quartz, as well as traces of Illite and calcite. Table 2 summarizes the physical characteristics of the dune sand.

II.1.2 Lime

The lime used in our study comes from Saida, which is characterized by a low concentration of oxide elements such as silicates SiO_2 and aluminates $A1_2O_3$ and a high concentration of basic elements such as free lime CaO. Its absolute density is between 2.2 and 2.4 g/cm3 and an apparent density (g/cm^3) as well as a specific surface area between 8000 and 12000 cm^2/g. Table 3 presents a summary of the chemical composition of the lime.

Table 3: Chemical composition of lime (%).

CaO	MgO	Fe_2O_3	AL_2O_3	SiO_2	SO_3	CO_2	$CaCO_3$
>73.3	<0.5	<2.0	<1.5	<2.5	<0.5	<5.0	<10.0

Source: Authors, (2026).

II.1.3 Cement

Chemical analysis of cement using a Scanning Electron Microscope (SEM) was carried out in the M'sila cement factory laboratory, the results are presented in Table 4.

Table 4: Chemical analysis of cement (%).

SiO_2	Al_2O_3	Fe_2O_2	CaO	MgO	SO_3	K_2O	Na_2O	Cl	CaO
19.70	4.52	3.49	62.15	1.79	2.27	0.49	0.25	0.02	1.00

Source: Authors, (2026).

II.2 METHODS

The research focuses on the impact of the incorporation of heavy substitutes, such as lime and cement, on the mechanical properties of the sand examined. The mixtures, which have been developed in this direction contains lime percentages of 4%, 6% and 8% and cement percentages of 5%, 7% and 9%. It is important to emphasize that these percentages were set in accordance with the guidelines of the standards concerning road networks [28] (GTR, 2000). The purpose of compaction is to densify the soil in place, to minimize its deformation and to optimize its bearing strength. Proctor standard compaction tests were carried out in accordance with the standard NF P94-093 [29] to determine the maximum dry density (MDD) and the optimum water content. Before being placed in the shear box, all mixtures underwent static compaction until an optimum Proctor was reached. The test is carried out at a rapid speed of 1 mm/min, which corresponds to an unconsolidated and undrained (UU) test in accordance with standard NF P 94-071-1[30]. It should be noted that the results regarding the optimum moisture content were also used to prepare the CBR and strength tests on the soil samples examined. Then, the CBR tests, shear strength, compressive strength and tensile strength were carried out in accordance with the standards NF P 94-078 [31], NF P 94-071-1 [30], standard EN 13286-53 [32] and standard EN 13286-42 [33], respectively, on the soil samples studied.

III. RESULTS AND DISCUSSIONS

III.1 EFFECT OF STABILIZERS ON THE PARAMETERS OF THE STANDARD PROCTOR COMPACTION TESTS

The obtained results are represented in Figures 2a and 2b. It is observable that the various additions seem to shift the coordinates of the Proctor modified (optimum moisture content and maximum dry density) to the right, by increasing the maximum dry density (MDD) and optimal water content (OMC) (Figures 4a and 4b). This trend of increasing the optimal water content becomes clearer for the case of cement mixtures. The densification of this material allowed us to note a dry density systematically higher than that of formulations based on cement. Lime mixtures result in an increase in dry density from 1.72 g/cm^3 to 1.81 g/cm^3. The optimum water content varies in the range of 5% to 7.8%. For cement-based formulations, the optimal water content varies in the range of 5% to 9.9% for a dry density between 1.72 g/cm^3 and 1.82 g/cm^3. The increase in density is attributed to the increase in the concentration of fine particles and their substitution by particles, which is the origin of the behavior observed in the various mixtures.

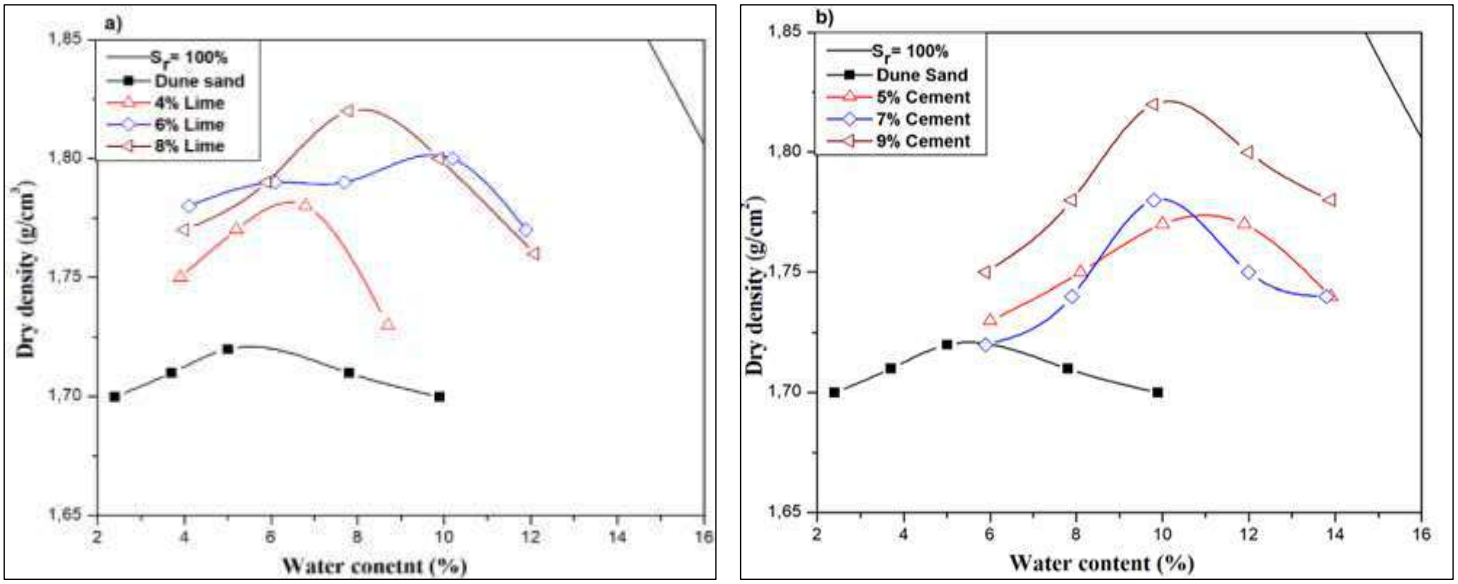


Figure 2: a) Modified Proctor of lime mixtures, b) Modified Proctor of cement mixtures.
Source: Authors, (2026).

III.2 EFFECT OF STABILIZERS ON THE CBR TESTS

To determine the CBR values at 95% of the optimum consists of punching into the CBR mold samples were made using various compaction energies (10 blows per layer, 25 c/c, 55 c/c), and at water content which corresponds to the optimal of the modified Proctor test. The aim being to obtain 3 castings with clearly different compactness (this compactness is respectively of the order of 90%, 95% and 100% of the maximum dry density of the modified Proctor). The material is then introduced according to the process of the modified Proctor test in 5 successive layers of equal thickness which will each be compacted with the same modified Proctor lady according to different compaction energies which are generally 10, 25 and 55 strokes per layer. According to the curves which represent the pressure of the piston as a function of its depression (Figure 3a and 3b), we determine the CBR index for each compaction energy.

This makes it possible to illustrate the change in dry density as a function of CBR index [γd = f(ICBR)] (Figure 4a and 4b). The final CBR index, determined graphically, is that which equals 95% of the maximum dry density of the modified Proctor test. Figure 5 a and b illustrates the variation of CBR at 95% for lime mixtures and cement mixtures. Figure5a show that lime improves the immediate bearing capacity from 8.28% to 13.45%for 4% lime. Beyond this value, we note a slight decrease to 12.32% for 8% lime. At the same time, in immersion, the lift improves from 6.45% to 12.10% for 6% lime then it drops to 9.5% for 8% lime. Just like lime, cement (Figure 5.b) showed a very significant increase in immersion lift of up to 162% in the case of 9% cement. This is obviously due to the cement setting. Unlike the lift in immersion, the immediate lift of cement shows values close to the values of lime where we note 26.5% for 5% cement then it drops to 21% for 9% cement.

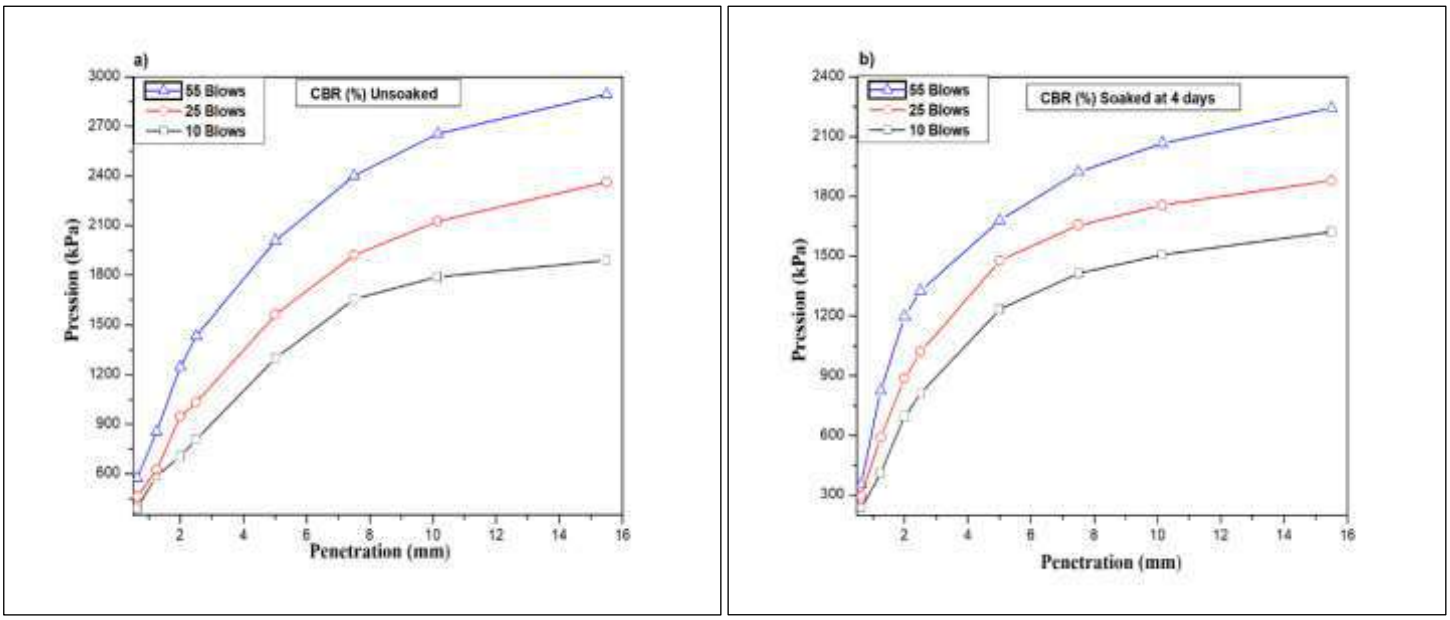


Figure 3: a) Variation of pressure according to the penetration (Unsoaked case), b) Variation of pressure according to the penetration (Soaked case).
Source: Authors, (2026).

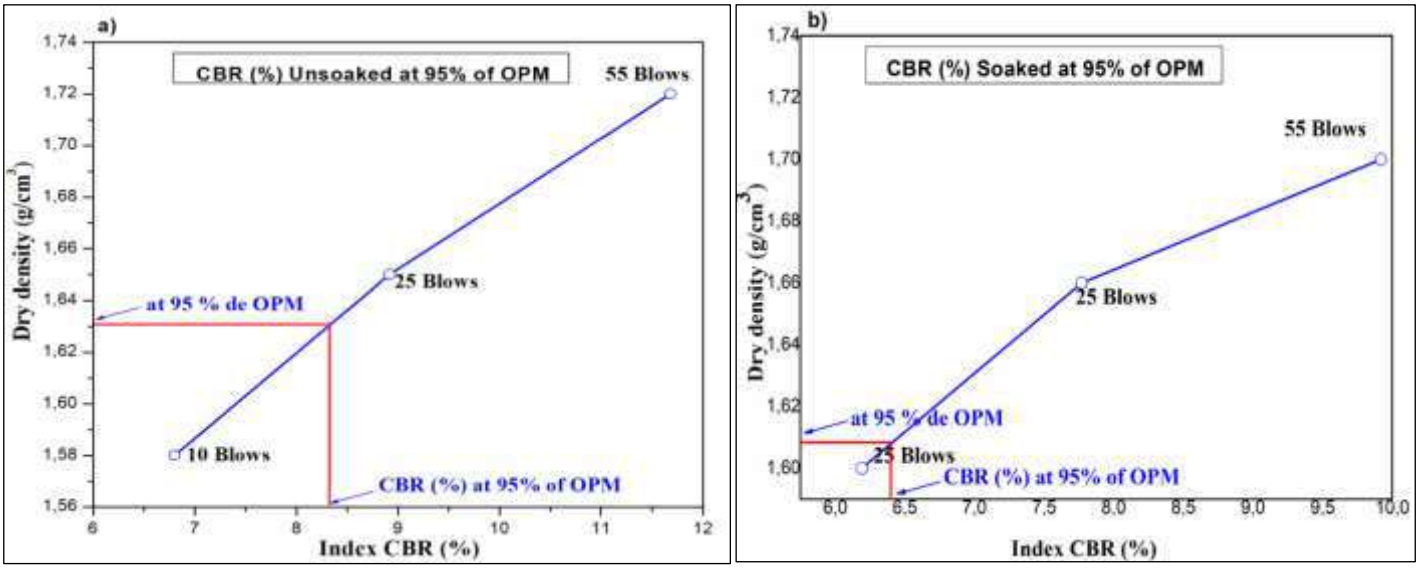


Figure 4: a) Variation of dry density according to the CBR Unsoaked index, b) Variation of dry density according to the CBR soaked index.
Source: Authors, (2026).

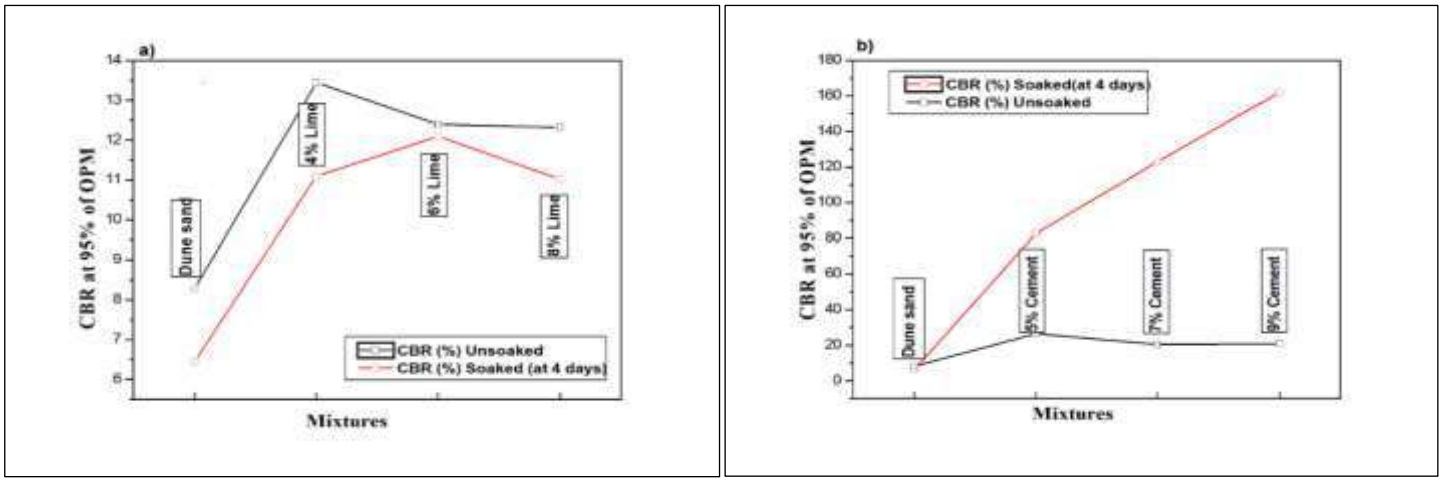


Figure 5: a) Variation of CBR index (unsoaked and soaked) according to the percentage oflime, b) Variation of CBR index (unsoaked and soaked according to the percentage of cement.
Source: Authors, (2026).

III.3 EFFECT OF STABILIZERS ON THE SHEAR STRESS TESTS

The test is intended to evaluate the mechanical characteristics of a soil, that is to say the cohesion C, the friction angle Φ and the shear resistance τ_{max} at the moment of rupture. The cohesion of a soil is the property that makes it possible to oppose the sliding of the grains which compose it and to resist a shear force, that is to say to oppose the sliding of a layer, this allowing exposure to the danger of landslides to be limited. The used test to determine the parameters is the unconsolidated undrained direct shear test (U.U) in accordance with standard NF P 94-071-1. The samples of the dune sand-lime, dune sand-cement mixtures were placed between two half-boxes, which have the ability to move relative to each other. In addition, a piston makes it possible to exert a stress normal to the shear plane. The samples of the sand-binder mixtures are sheared at their compaction water contents. The standard constraints applied during these tests are as follows:

σ1 = 100 kPa; σ2 = 200 kPa; σ3 = 300 kPa. The shear speed used is 1mm/min.

The mechanical parameters of the mixtures are determined by graphically representing the Coulomb lines (intrinsic curves) in an orthonormal coordinate system, where the abscissa axis indicates the vertical stresses (σi) and the ordinate axis the maximum shear stresses (τimax). The slopes of these lines represent the tangents of the internal friction angles Φ of the sand-binder mixtures, on the one hand. On the other hand, the cohesions C are obtained by the intersection of these lines with the ordinate axis. The values of the angle of friction and cohesion are obtained analytically according to Coulomb's law:

$$\tau = C + \sigma\, tg(\Phi) \quad (1)$$

Figure 6a and 6b show the intrinsic curves of the dune sand-lime and dune sand-cement mixtures. Figure 7 and 8 present the variations of the cohesion of the mixtures and the variation of the friction angle depending on the binders added. The dune sand-lime and dune sand-cement mixtures show an increase in cohesion ranging from 2 kPa to 81 kPa for 6% lime and 46 kPa for 5% cement. Beyond these contents, we observe a slight decrease in cohesion, which becomes around 24 kPa for 8% lime and 15 kPa for 9% cement. However, the function of cement is to strengthen the structure of the sand by binding the particles together, thanks to the setting of the binder; on the other hand, the dune sand-cement mixture becomes less flexible and risks cracking quickly under traffic. However, it is necessary to choose a proportion of cement which improves the bearing capacity without making the mixture too stiff once it has hardened. Unlike cohesion, the friction angle shows a decrease to 32° for 5% cement and to 21° for 6% lime followed by an increase to 40° for 9% cement and to 27° for 8% lime. It is evident that all the dosages of hydraulic binders analysed demonstrate an increase in cohesion which is inversely proportional to the friction angle.

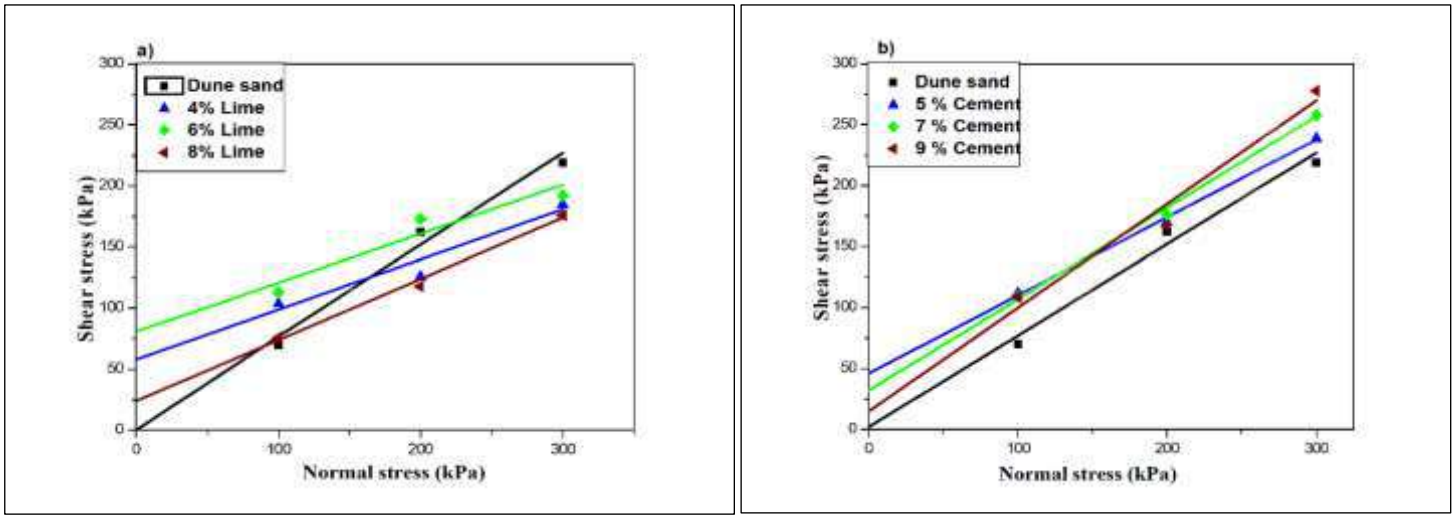


Figure 6: a) Intrinsic curves of lime mixtures, b) Intrinsic curves of lime mixtures.
Source: Authors, (2026).

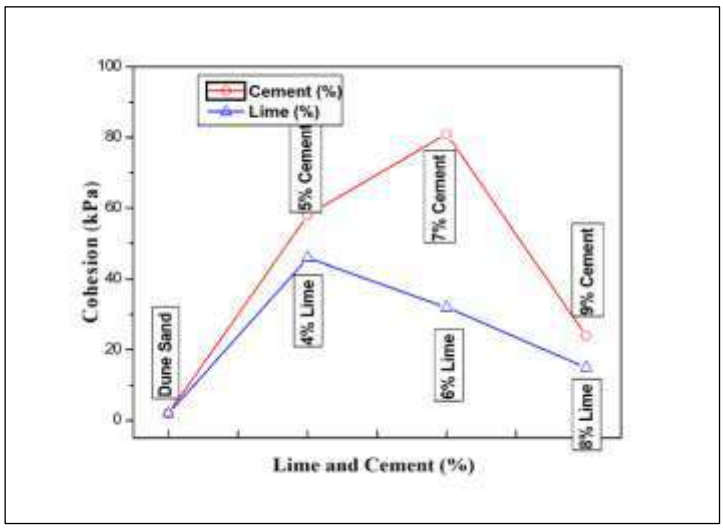


Figure 7: Evolution of the cohesion according to the binders added.
Source: Authors, (2026).

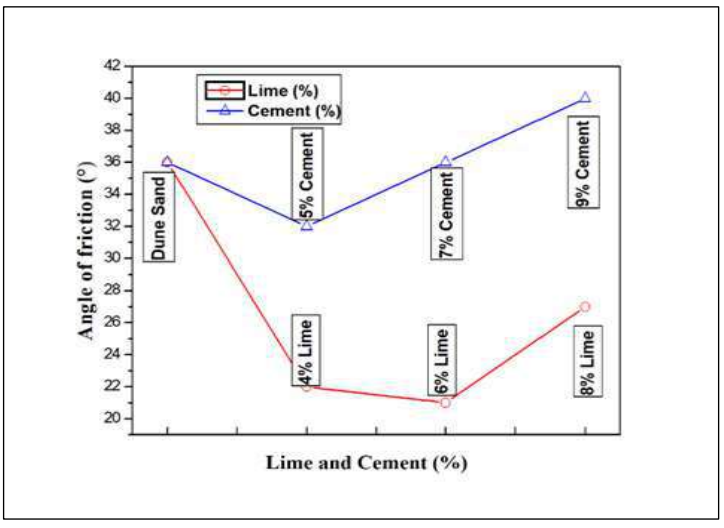


Figure 8: Evolution of the angle of frictionaccording to the binders added.
Source: Authors, (2026).

III.4 EFFECT OF STABILIZERS ON THE COMPRESSIVE AND TENSILE STRENGTH TEST

In accordance with standard 13286-53, the series of cylinders with dimensions φ = 100 mm and H = 100 mm was set up for the various mixtures. The mixture was arranged in a uniform layer in a double piston mold and then statically compressed until the optimum dry density and moisture content were reached, established according to the modified Proctor method in accordance with EN 13286-41. Following their preparation in accordance with standard EN 13286-53[32], the samples were kept in bags at 20 ± 2 °C until testing. The different mixtures were tested on days 7, 14, 28, 60, and 90. The tensile tests were performed on cylindrical samples for the different mixtures. The tensile tests were carried out on cylindrical specimens of the same dimensions in accordance with standard EN 13286-42 [33] by measuring the indirect tensile strength, but the specimens were compacted and preserved in the same way until the date of the test (at 90 days). Figures. 9a and 9b show the progression of compressive strength over various curing times and binder ratios. The compressive strength capacity following curing time demonstrates a moderate growth up to 265.59 kPa for a curing duration of 90 days (Figure 9a) for 8% lime mixture.

It is clear from the results that lime does not act as a good stabilizer on sandy material since only the role of grain size corrector is predominant. For the case of cement (Figure 9b), the results are greater than those obtained with lime and go up to 2734.25 kPa and 2973.41 kPa respectively for mixtures with 7% and 9% cement for a curing duration of 28 days and at 3425.26 kPa and 3845.68 kPa for a cure duration of 90 days. This lets us notice that the increase in compressive strength is less significant from 28 days to 90 days, which explains the rapid hydration of the cement for the first period (28 days). Figures 10a and 10b illustrate the variation of the tensile strength depending on the binders incorporated. The tensile strength generally increases with the percentage of binder added. According to Figure 10a, lime has low values before 8% lime, value for which the resistance becomes more significant. For the case of cement, we notice an increase in the tensile strength up to the respective values of 398.23 kPa and 538.75 kPa corresponding to the percentages of 7% and 9% respectively for a curing duration of 28 days. It is clear that cement mixtures give better tensile strength compared to those obtained with lime mixtures.

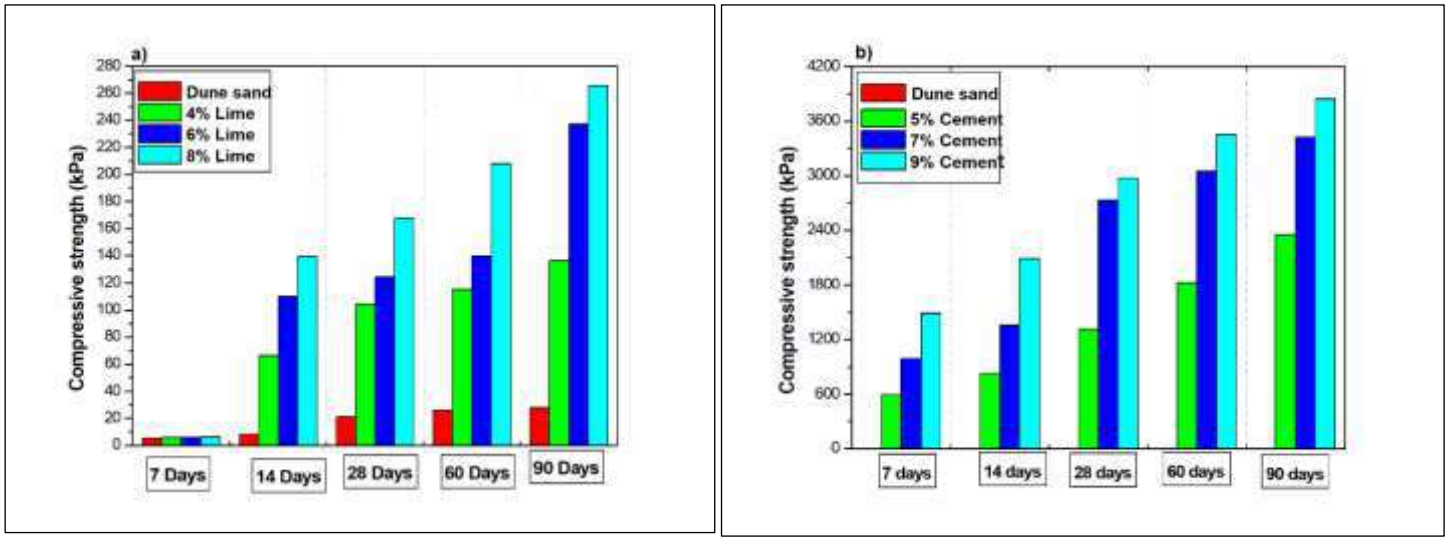


Figure 9: a) Evolution of the compressive strength according to the percentage of lime, b) Evolution of the compressive strength according to the percentage of cement.
Source: Authors, (2026).

In order to valorize dune sand, taking into account the environmental and economic aspects, the work focused on materials formulation techniques by adding hydraulic binders: lime and cement. The aims set were to examine the impact of the incorporation of percentages of fines (hydraulic binders) on the improvement of the mechanical characteristics of the dune sand treated. The results obtained from the present study show that the use of hydraulic binders and lime acts, in a very significant way, on improving the mechanical properties of this treated material. The following main conclusions can be drawn:

- Hydraulic binders such as lime and cement have a very positive effect on the mechanical performance of dune sand-based mixtures for use as pavement bases.
- Lime is less effective in improving mechanical characteristics than cement.
- Lime as a stabilizer is more suitable for fine soils (silt and clay) than for grainy soils (sand). It acts as a grain size corrector rather than a stabilizer.
- The addition of increasing quantities of hydraulic binder demonstrated a trend towards a permanent increase in the optimal water content.
- The use of hydraulic binders improves the tested mechanical performance (tensile and compressive strength).

IV.AUTHOR'S CONTRIBUTION

Conceptualization: Ali Smaida, Dif Fodil, Khalil Latreche, Smail haddadi, Hayet Cherfa.
Methodology: Ali Smaida, Dif Fodil, Khalil Latreche, Smail haddadi, Hayet Cherfa.
Investigation: Ali Smaida, Dif Fodil, Khalil Latreche, Smail haddadi, Hayet Cherfa.
Discussion of results: Ali Smaida, Dif Fodil, Khalil Latreche, Smail haddadi, Hayet Cherfa.
Writing – Original Draft: Ali Smaida, Dif Fodil, Khalil Latreche, Smail haddadi, Hayet Cherfa.
Writing – Review and Editing: Ali Smaida, Dif Fodil, Khalil Latreche, Smail haddadi, Hayet Cherfa.
Resources: Ali Smaida, Dif Fodil, Khalil Latreche, Smail haddadi, Hayet Cherfa.
Supervision: Ali Smaida, Dif Fodil, Khalil Latreche, Smail haddadi, Hayet Cherfa.
Approval of the final text: Ali Smaida, Dif Fodil, Khalil Latreche, Smail haddadi, Hayet Cherfa.

V ACKNOWLEDGMENTS

This research was funded by the laboratory of Laboratory Environment, Water, Geomechanics and Structures, (LEEGO), USTHB, Algeria and Civil Engineering and Sustainable Development Laboratory (LGCDD), University of Djelfa, Algeria.

VI. REFERENCES

[1] Rammal MM, Jubair AA. Sand dunes stabilization using silica gel and cement kiln dust. 2015; 18:179–91.

[2] Bell FG. 2005. Engineering Treatment of Soils. London and New York: Taylor & FrancisGroup.

[3] Rasul. J.M, Burrow M.P.N., Ghataora G.S. Consideration of the deterioration of stabilised subgrade soils in analytical road pavement design, Transportation Geotechnics. 2016, doi: http://dx.doi.org/10.1016/j.trgeo.2016.08.002.

[4] Talal S Amhadi, Gabriel J Assaf. Assessment of strength development of cemented desert soil. International Journal of Low-Carbon Technologies, Volume 14, Issue 4, December 2019, Pages 543–549, https://doi.org/10.1093/ijlct/ctz047

[5] Panwar K, Ameta NK. Open access stabilization of fine sand with ceramic tiles waste as admixture for construction of embankment. American Journal of Engineering Research (AJER), 2016; 5:206–12.

[6] Almadwi Fathi S, Assaf Gabriel J. Finding an optimal bitumen and natural sand balance for hot mix asphalt concrete in hot and arid regions. In: Proceedings of the 5th GeoChina International Conference—Civil Infrastructures Confronting Severe Weathers and Climate Changes in HangZhou, China, 2018, Springer: Cham pp. 1–12.

[7] Rou (2004), Le traitement des sols à la chaux et/ou aux liants hydrauliques pour l'exécution des remblais et des couches de forme, l'industrie cimentière française edn.

[8] Giummarra GJ. Unsealed roads manual: guidelines to good practice. 3rd ed. Vermont South: ARRB Group; 2009.

[9] A Gueddim, N Bouarissa, A Villesuzanne. Pressure dependence of elastic constants and related parameters for rocksalt MgO Computational materials science 48 (3) (2010), 490-494, https://doi.org/10.1016/j.commatsci.2010.02.010

[10] N Drissi, A Gueddim, N Bouarissa. First-principles study of rocksalt $Mg_xZn1_{-x}O$: band structure and optical spectra Philosophical Magazine 100 (12) (2020) 1620-1635, https://doi.org/10.1080/14786435.2020.172797

[11] A Gueddim, N Bouarissa. Theoretical investigation of the conduction and valence band offsets of GaAs1− x Nx/GaAs1− yNy heterointerfaces. Applied surface science 253 (17) (2007) 7336-7341, https://doi.org/10.1016/j.apsusc.2007.03.019

[12] A Amara, L Gacem, A Gueddim, R Belbal, MT Soltani, L Guerbous. Luminescence properties of Cr3+ ions in Na2ZnP2O7 crystal Physica B : Condensed Matter 545 (2018) 408-412, https://doi.org/10.1016/j.physb.2018.07.008.

[13] F Saidi, IS Yahia, HY Zahran, H Algarni. Correlations between Thermal Parameters and Glass Forming Abilities of Bulk Metallic Glasses Unveiled by Datamining ECS Journal of Solid State Science and Technology 11 (9) (2022) 091004, DOI 10.1149/2162-8777/ac8ba6

[14] K Benyelloul, A Djellouli, H Aourag, K Khodja, Y Bouhadda, A Adjaj. A principal component analysis of complex Iridium hydrides $XYIrH_{6(}$ X=K, Na: Y= Ba, Sr, Ca) and Osmium hydrides $ZOsH_6$ (Z= Ba, Sr, Ca, Mg) 2020 6th International Symposium on New and Renewable Energy (SIENR) (2020) 1-4,DOI : 10.1109/SIENR50924.2021.9631893

[15] K Benyelloul, L Seddik, Y Bouhadda, M Bououdina, H Aourag, K Khodja Effect of pressure on structural, elastic and mechanical properties of transition metal hydrides Mg7TMH16 (TM= Sc, Ti, V, Y, Zr and Nb): first-principles investigation Journal of Physics and Chemistry of Solids 111 (2017) 229-237, https://doi.org/10.1016/j.jpcs.2017.08.001

[16] H Faraoun, H Aourag, C Esling, JL Seichepine, C Coddet. Elastic properties of binary NiAl, NiCr and AlCr and ternary Ni2AlCr alloys from molecular dynamic and Abinitio simulation. Computational materials science 33 (1-3) (2005) 184-191, https://doi.org/10.1016/j.commatsci.2004.12.011

[17] Smaida, A., Haddadi, S., Nechnech, A. Improvement of the mechanical performance of dune sand for using in flexible pavements. Constr Build Mater. 2019. 208. Pp. 464–471. https://doi.org/10.1016/j.conbuildmat.2019.03.041

[18] Smaida, A., Cheriet , R., Haddadi , S., & Nechnech , A. (2022). Valorization of dune sand of Djelfa (Algeria) treated by hydraulic binders in pavement foundations. Journal of Building Materials and Structures, 9(1), 33-43. https://doi.org/10.34118/jbms.v9i1.1897

[19] Ali Smaida, Mustapha Cheddad, Abderrahman Hamadi, Koribi Hadi, Mohamed Ben Slim, Hayet Cherfa, Amina Bouguerroua, Smail Haddadi. Stabilization of dune sand using crushed sand, silt and blast furnace slag for use in road field. Construction and Building Materials 406 (2023) 133397. https://doi.org/10.1016/j.conbuildmat.2023.133397

[20] Ismael. Nabil F. Influence of fines on properties of arid climate sand deposits, fourth International Conference on Unsaturated Soils held in Carefree, Arizona, April 2-6, 2006; https://doi.org/10.1061/40802(189)135

[21] Ben Dhia. M.H (1998) : Quelques particularités de l'utilisation du sable de dune en construction routière en milieu saharien, bulletin des laboratoires des ponts et chaussées, P 33- 42.

[22] Baghdadi. Z.A., Rahman M. A. (1990) : The Potential of Cement Kiln Dust for the Stabilization of Dune Sand in Highway Construction; Building and Environment, Vol. 25, No. 4, pp. 285-289. https://doi.org/10.1016/0360-1323(90)90001-8

[23] Ghrieb A, Kettab R. Stabilization and Utilization of Dune Sand in Road Engineering Arab science journal engineering, 2014 39:1517–1529 DOI 10.1007/s13369-013-0721-z.

[24] Cherfa, H., Nechnech, A., Saoudi, N., Ait Mokhtar, K. Mechanical properties of slag sand mixture used in road pavements. Magazine of Civil Engineering. 2021. 108(8). Article No. 10806. DOI: 10.34910/MCE.108.6

[25] ElhadjGuesmia Daheur, ZhongSen Li, Abdellah Demdoum, Said Taibi, Idriss Goual. Valorisation of dune sand-tuff for Saharan pavement design. https://doi.org/10.1016/j.conbuildmat.2022.130239

[26] Venkatarama. R. B.V, AJAY. G. Influence of sand grading on the characteristics of mortars and soil–cement block masonry, Construction and Building Materials, 22: 1614– 1623, 2008. https://doi.org/10.1016/j.conbuildmat.2007.06.014

[27] Haach V.G, Vasconcelos G, Lourenço P.B, Influence of aggregates grading and water/cement ratio in workability and hardened properties of mortars, Construction and
Building Materials 25: 2980–2987, 2011. https://doi.org/10.1016/j.conbuildmat.2010.11.011

[28] GTR. (2000) Guide technique pour la réalisation des remblais et des couches de forme. Editions du SETRA-LCPC, Fascicules I & II, 98 p. et 102 p.

[29] AFNOR 1999, NF P94-093. Sols : reconnaissance et essais - Détermination des références de compactage d'un matériau - Essai Proctor normal. Essai Proctor modifié.

[30] AFNOR 1994, NF P94-071-1. Sols ; reconnaissance et essais – essais de cisaillement rectiligne à la boîte partie 1 : Cisaillement direct.

[31] AFNOR, NF P94-078. Sols : reconnaissance et essais – Indice CBR après Immersion-Indice CBR immédiat. (1997).

[32] AFNOR NF EN 13286-53. Mélanges traités et mélanges non traités aux liants hydrauliques – Partie 53 : méthode de confection par compression axiale des éprouvettes de matériaux traités aux liants hydrauliques. 2005.

[33] AFNOR, NF EN 13286-42. Mélanges traités et mélanges non traités aux liants hydrauliques – Partie 41 : méthode d’essai pour la détermination de la résistance à la compression des mélanges traités aux liants hydrauliques. 2003.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p.705-717. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3032

RESEARCH ARTICLE OPEN ACCESS

DC-BUS CAPACITOR DISCHARGE CONTROL STRATEGY FOR PMSM DRIVES SYSTEM: ELECTRIC VEHICLES

Chevuru Mahima*[1], Ch. Chengaiah[2]

[1]Department of Power Systems Engineering, India.
[2]Sri Venkateswara University College of Engineering, S V University, Tirupati, India. 517502.

[1]https://orcid.org/0009-0003-0929-7074, [2]https://orcid.org/0000-0002-4604-9497

Email: *chevurumahima@gmail.com, chinthapudisvu@gmail.com

ARTICLE INFO

Article History
Received: December 1, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Permanent Magnet Synchronous Machine (PMSM),
DC- bus discharge,
Extended state observer (ESO), ANN,
Adaptive Neuro-Fuzzy Inference System (ANFIS) etc.

ABSTRACT

Electric vehicles have become increasingly popular for their high efficiency and zero emissions, making them vital for sustainable transportation. The Permanent Magnet Synchronous Motors (PMSM) are widely used in Electric Vehicles, because of the Active discharge mechanism play a critical role in maintaining the safety and reliability of Electric Vehicle (EV) system. The proposed control strategy ensures that, the DC bus capacitor voltage is safely discharged to an acceptable level during emergency fault conditions. This paper proposed a robust winding-based capacitor discharge strategy that effectively utilizes the inherent machine windings of the PMSM as an alternative discharge path when the active circuit is unavailable. The control strategy discharge mechanism operates in two distinct stages to ensure rapid and controlled voltage reduction. In the first stage, a large flux-enfeebling in d-axis current which is injected to quickly suppress the back electromotive force (EMF), thereby achieving an immediate reduction in the DC bus voltage. In the second stage, a voltage regulation process compensates for residual energy caused by switching losses and inductive effects, maintaining the voltage at a safe level until the system stabilizes. This approach eliminates the dependence on external resistors for improving the system safety and cost-effectiveness. During normal operation to enhance the dynamic performance of the electric vehicle system, a PMSM drive systems with an integrated ANFIS controller is used for speed regulation. The proposed control strategy synergizes the adaptive learning capability of neural networks with the rule-based reasoning of fuzzy logic, to resulting in improved nonlinear control characteristics compared with PI, FUZZY and ANN Controllers. This paper presents the design and simulation of an PMSM drive system with ANFIS based speed control for EV system. To validate the effectiveness of the proposed system, the MATLAB Simulink model is implemented on a three-phase Surface-mounted PMSM (SPMSM) drive platform. The PMSM drive with ANFIS-controller, exhibits precise and stable speed control, even under rapid transients and varying load conditions and enhanced performance during standard operation offering a reliable, intelligent and cost-efficient solution for modern electric vehicle drive systems.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The permanent magnet synchronous motor (PMSM) is commonly used as the power train in electric vehicles [1-4] due to on hand benefits, such as high efficiency, high torque density, and excellent torque ratios [5-9]. Additionally to enhance the driving efficiency of PMSMs and minimize the driving currents increasingly rely on high-voltage drive systems [10]. However, high-voltage power train pose a risk of electric shock to the driver in the event of an emergency, such as a car crash, to address this concern, the United Nations Vehicle Regulation ECC R94 mandates that the DC bus capacitor voltage is reduced to safe level with a short duration [11], [12].

The winding-based discharge method can not only reduce the volume of the discharge circuit but also be cost- effective [13-15]. The schematic block diagram of the EV powertrain actuation system as shown in Figure 1. The topology of an Electric Vehicle (EV) powertrain employing a Permanent Magnet Synchronous Motor (PMSM). The system includes a battery pack, DC–DC boost converter, inverter, PMSM, and gearbox. The battery supplies DC power, which is boosted and converted into three-phase AC by the inverter to drive the PMSM. The motor output is transmitted to the wheels through the gearbox.

A breaker is incorporated for safety and is tripped during emergencies or crashes to isolate the battery, ensuring system protection and operational safety. A dc-bus voltage regulator based on a proportional integral controller and a modulation index controller are used to generate the reference d-axis current in order to stop the bus voltage. A fixed d-axis current and piecewise q-axis current control is suggested in [16] with the goal of discharging the capacitor voltage of specific EV driven systems with large inertia and minimal safe current. Furthermore, the synchronous reference frame of the three-phase PWM rectifier model is nonlinear, making the ANFIS based approach inapplicable to the system. The purpose of this study is to provide a reliable dc-bus capacitor discharge method for PMSM EV drives in the event of a discharge circuit failure. To achieve the dc-bus voltage constant and prevent voltage spikes, the measured total disturbance is instantly corrected to the dc-bus capacitor energy loop control law.

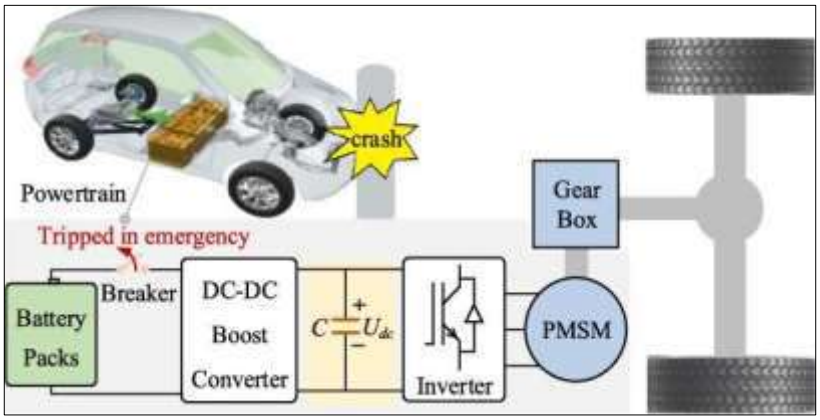


Figure1: Schematic diagram of PMSM-EV powertrain systems.
Source: Authors, (2026).

In the event of an emergency involving electric vehicles, a malfunctioning or failed bleeding resistor can pose serious risks of electric shock to both passengers and rescuers. To enhance EV safety, a practical solution is to implement advanced discharge circuit fault-handling techniques. These techniques ensure that the discharge process adheres to critical safety standards, such as completing discharge within five seconds while preventing voltage surges. By integrating these safeguards, the risk of high-voltage exposure can be significantly minimized, making emergency responses safer and more effective. To stabilize the voltage, ESO is used to monitor the dc-bus capacitor energy until the voltage drops to zero. The active disturbance rejection observer, which includes ESO [17]. The mathematical model of the Permanent Magnet Synchronous Motor (PMSM) in the rotating dq-reference frame is employed because it accurately represents both the electrical and mechanical dynamics of the motor during the capacitor discharge operation. The following equations describe the stator voltage, current variation, and torque production in the PMSM [18-20].

$$\begin{cases} u_d = R_s i_d + L_s \dfrac{di_d}{dt} - \omega_e L_s i_q \\ u_q = R_s i_q + L_s \dfrac{di_q}{dt} + \omega_e L_s i_d + \omega_e \psi_f \\ T_e = \dfrac{3}{2} p \psi_f i_q \end{cases} \tag{1}$$

Where i_d, i_q and u_d, u_q: d axis and q-axis of stator currents and voltages, L_s: stator inductance, R_s: Stator resistance, ω_e: electrical angular velocity, p: pole pair, ψ_f: flux linkage T_e: electromagnetic torque. Under steady state condition the mechanical equation of the permanent magnet synchronous motor can be expressed as,

$$J \frac{d\omega_m}{dt} = T_e - B_m \omega_m \tag{2}$$

Where ω_m : mechanical angular velocity, J and B_m: inertia and mechanical damping coefficient. Under dynamic state condition the dc capacitor voltage u_{dc} can be expressed based on the 3-phase inverter circuit which is shown in Figure 2

$$C \frac{du_{dc}}{dt} = i_{dc} \tag{3}$$

Where i_{dc} and u_{dc} stand for the dc-bus current and voltage, respectively, and C is the capacitor.

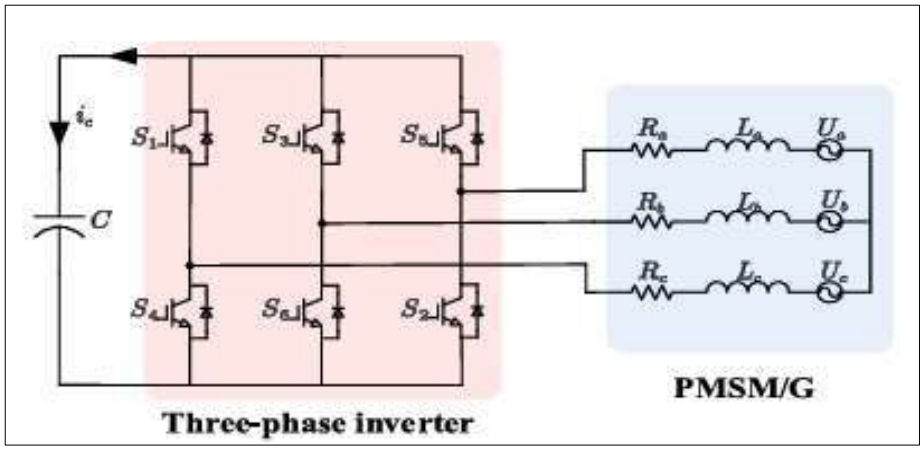


Figure 2: Three phase inverter with PMSM winding based dc-bus capacitor.
Source: Authors, (2026).

The Figure 2 shows that the three phase inverter with PMSM winding based dc bus capacitor discharge, however from the basic fundamentals the converter's power loss, the AC side's power should match the DC side's power, which means

$$u_{dc}i_{dc} + P_{con} = -\frac{3}{2}\left(u_d i_d + u_q i_q\right) \tag{4}$$

Where p_{con}: Power switching loss. To examine the power flow process during discharge, substitute (1) and (3) equations into the instantaneous power balance equation (4).

$$\frac{d}{dt}\left(\frac{1}{2}Cu_{dc}^2\right) = -\frac{3}{2}\omega_e\psi_f i_q - \frac{3}{2}R_s\left(i_d^2 + i_q^2\right) - \frac{3}{2}\left(L_s i_q\frac{di_q}{dt} + L_s i_d\frac{di_d}{dt}\right) - P_{con} \tag{5}$$

From equation. (5) it is observed that, a portion of the permanent magnet synchronous motor's electromagnetic power output is converted to inductance coil energy, copper winding loss, and converter loss. The capacitor's energy storage is then modified using the residual.

II. PROPOSED MODEL OF THE ROBUST WINDING BASED DISCHARGE METHOD

When EVs experience an emergency with an active discharge circuit defect, it's critical to redesign a robust discharge strategy in order to achieve a safe and dependable discharge technique. Thus, this section presents a reliable winding-based dc-bus capacitor discharge method based on ESO.

II.1 RAPID DISCHARGE STATE:

When an active discharge is needed, it is significant and safe for both passengers and rescuers if the dc-bus voltage can drop to safe levels right away. When an emergency occurs in an electric car, the back electromotive force (EMF) is more than the safe voltage because the motor speed is higher. Due to the dc-bus capacitor voltage follow the back EMF, the dc-bus capacitor voltage can drop to safe voltage when the back EMF is reduced to safe voltage. According to the flux-enfeebling principle, a large d-axis current is injected to rapidly reduce the EMF because the dc-bus voltage is proportional to the PMSM's EMF. At the same time, the q-axis current remains at zero to stop the conversion of kinetic energy into electrical energy [21]. From EVs PMSM powertrain system which is shown in fig.1. The dc-bus voltage is expressed as

$$u_{dc} = \sqrt{3}\omega_e\left(\psi_f + L_s i_d\right) - R_s\sqrt{i_d^2 + i_q^2} - 2V_{on} \tag{6}$$

Where V_{on} : diode conducted voltage. The voltage should decrease to a safe level during the rapid discharge stage. Then, the flux-enfeebling current on the d-axis is expresse

$$i_d^* = \frac{U_{safe} - \sqrt{3}\omega_e\psi_f}{\left(\sqrt{3}\omega_e L_s + R_s\right)} \tag{7}$$

Where U_{safe} : safe voltage 60V. Consequently, once the active discharge is needed, the estimated d-axis current can be injected to lower the dc-bus voltage to the safe voltage. A zero q-axis current is applied concurrently to stop any electromagnetic power from turning into electrical energy.

II.2 BUS VOLTAGE CONTROL STAGE ON ESO

The Extended State Observer (ESO) is a powerful control method widely used to manage disturbances and enhance the robustness of dynamic systems, particularly in challenging environments in electric vehicle (EV) powertrains. In the context of Permanent Magnet Synchronous Motor (PMSM) drives, the ESO plays a crucial role in managing uncertainties, compensating for system disturbances, and improving control precision. When the DC-bus voltage reaches 60 V, the ESO-based voltage control stage is activated. To maintain the voltage within a safe range 60 V the controller regulates the discharge process and ensures stable operation, a robust voltage control loop is designed. Since the energy flow during the discharge mode described in Equation (5) exhibits nonlinear characteristics, the PI, FUZZY controller is unable to produce adequate results, particularly when motor parameters will alter during the discharging mode to overcome this ANFIS with ESO-based ultra-local model is introduced to effectively address this issue. For converting Equation (5) into an integral series form, the first-order ultra-local model of a single-input and single-output system is defined as given in [22–25].

$$y^{\cdot} = F + \alpha u \tag{8}$$

Where F Stands for known and unknown components of the system, u : control variables, y : output variables, and α : nonphysical scaling factor. This article's mathematical model of the energy flow-based ESO can be stated using formulas (5) and (8).

$$\frac{dE_c}{dt} = F + \alpha u \tag{9}$$

Where E_c: $0.5Cu^2_{dc}$, u: i_q, α: $-1.5\omega_e\psi_f$ and the other terms on the right side of equation (5) except $-1.5\omega_e\psi_f i_q$ are treated as the total disturbance and notated as F

$$F = -\frac{3}{2}R_s\left(i_d^2 + i_q^2\right) - \frac{3}{2}\left(L_s i_q \frac{di_q}{dt} + L_s i_d \frac{di_d}{dt}\right) - P_{con} \tag{10}$$

Equation (10), shows that the total disturbance corresponds to the overall power loss, which includes switching loss, winding copper loss, and inductive coil energy storage. It is assumed that *F* remains constant during each sampling period. A linear equation of ESO, can be formulated based on the ultra-local model, which is shown in fig.4 proportionately considering E_c and F as a variables given in Equation (8) is expressed as.

$$\begin{cases} e_1 = z_1 - E_c \\ \dot{z}_1 = z_2 - \beta_1 e_1 + \alpha u \\ \dot{z}_2 = -\beta_2 e_2 \end{cases} \tag{11}$$

Where z_1 and z_2: estimated value of E_c and F. β_1 and β_2: feedback gains it can be estimate the real-time value of E_c and F that is, $Z_1 \rightarrow E_C$; $Z_2 \rightarrow F$. From the above mathematical model analysis the winding based DC- bus capacitor discharge is implemented with proposed control strategy mechanism is described with ANFIS and ESO which is described in the subsequent sections.

III. WINDING BASED DC BUS CAPACITOR DISCHARGE STRATEGY WITH ANFIS

The conventional and suggested discharge methods using the dc-bus voltage controller based on ESO. It also presents the suggested resilient dc- bus capacitor discharge strategy. The proposed winding based dc-bus capacitor discharge control system as shown in fig.3 and the corresponding bus voltage control strategy on ESO of internal diagram as shown in fig.4. The proposed strategy consists of two stages. In the rapid discharge stage, a large flux-enfeebling d-axis current is injected into the motor, reducing the back electromotive force (EMF) and consequently lowering the dc-bus voltage to a predefined safe threshold. During this stage, the q-axis current is maintained at zero to prevent additional energy conversion from mechanical to electrical form.

Once the safe voltage is reached, the process transitions into the voltage regulation stage. During normal driving, Port-1 is turned ON the motor uses vector control to run the vehicle, while the ESO helps the current follow the reference smoothly. When the system receives a discharge command, Port-2 is turned ON, and the ESO-based discharge control is used to safely and quickly reduce the dc-bus capacitor voltage through the motor windings [26]. ANFIS are used in the d-axis and q-axis current loops so that the current can respond fast and accurately in both normal operation and discharge mode. Where as ESO is employed to estimate and compensate for system disturbances such as copper losses, switching losses and inductive energy storage

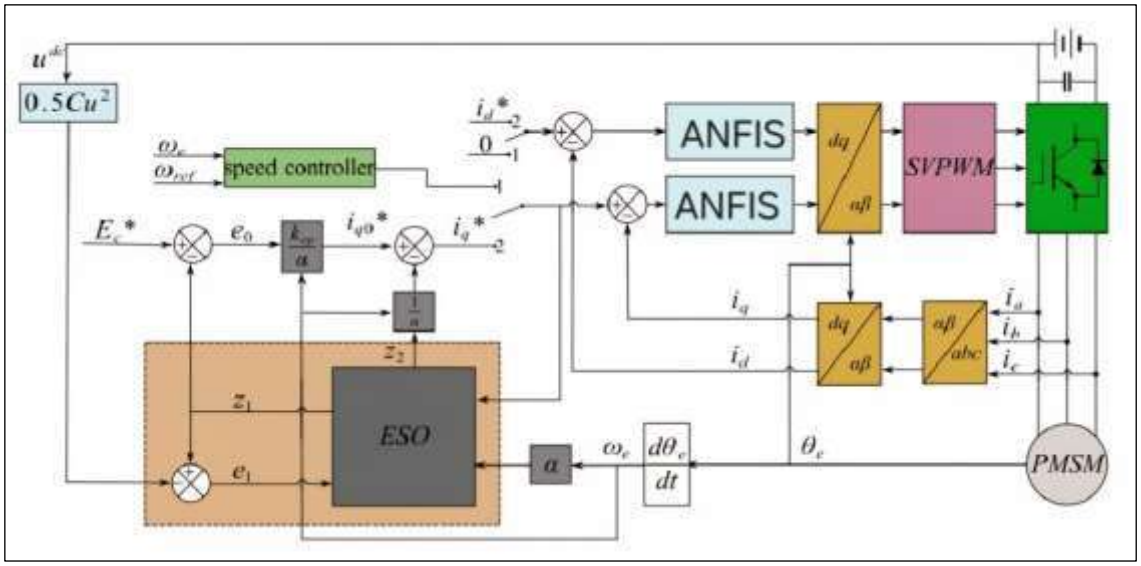


Figure 3: Block diagram of the winding based dc bus capacitor discharge strategy.
Source: Authors, (2026).

The ESO shown in the above block diagram fig.3 is represented as a subsystem in the fig.4 below, showing its internal structure.

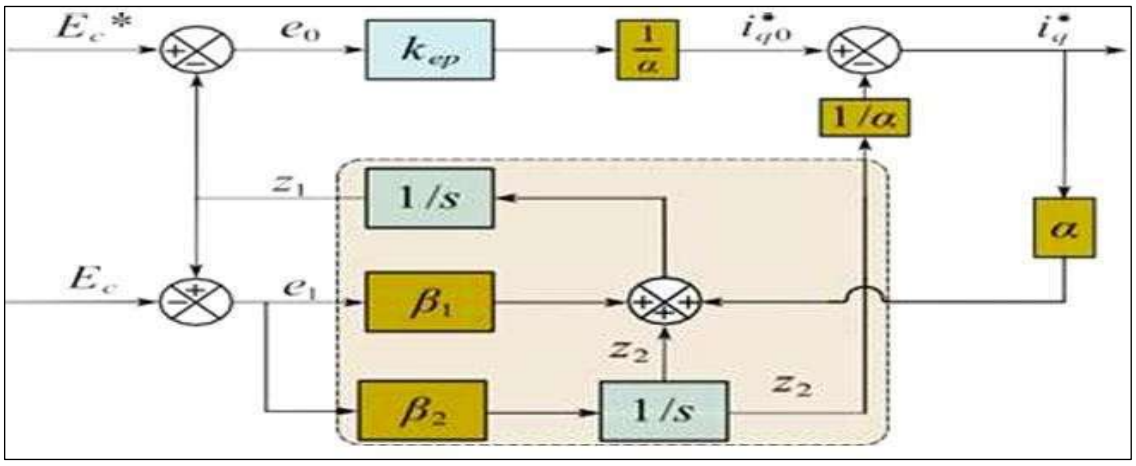


Figure 4: Bus voltage control strategy on Extended State Observer (ESO).
Source: Authors, (2026).

The ESO-based design enhances the robustness of the discharge control by effectively addressing nonlinear dynamics and parameter variations that conventional PI controllers cannot handle. In addition to ensuring safety under emergency conditions, maintaining high performance under normal operating modes is equally important for EV traction systems [27]. This work also incorporates an Adaptive Neuro-Fuzzy Inference System (ANFIS) controller for PMSM speed regulation. The ANFIS controller combines the self-learning capability of neural networks with the reasoning ability of fuzzy logic. This hybrid intelligent control strategy improves dynamic response, robustness, and overall stability of the PMSM drive during normal operation.

IV. PROPOSED ADAPTIVE NEURO FUZZY INFERENCE SYSTEM (ANFIS) CONTROLLER

Fig.5 shows that an Adaptive Neuro-Fuzzy Inference System (ANFIS) consists of a five-layer feed-forward structure that integrates the human-interpretable fuzzy logic framework with the adaptive learning capability of neural networks. Each layer performs a specific computational function in the Sugeno-type fuzzy inference process. The rules are expressed which is shown in table.1

Table:1 Sugeno ANFIS Rule Base.

Rule	Antecedent (IF …)	Consequent (THEN …)
R1	If (x) is (A_1)	($f_1 = p_1x + q_1$)
R2	If (x) is (A_2)	($f_2 = p_2x + q_2$)
R3	If (x) is (A_3)	($f_3 = p_3x + q_3$)

Source: Authors, (2026).

Here, $\boldsymbol{A_1}, \boldsymbol{A_2}$,and $\boldsymbol{A_3}$ denote the three membership functions corresponding to the input variable. The parameters $\boldsymbol{p_i}$ and $\boldsymbol{q_i}$ represent the consequent parameters, which are tuned during the learning process to achieve optimal system performance.

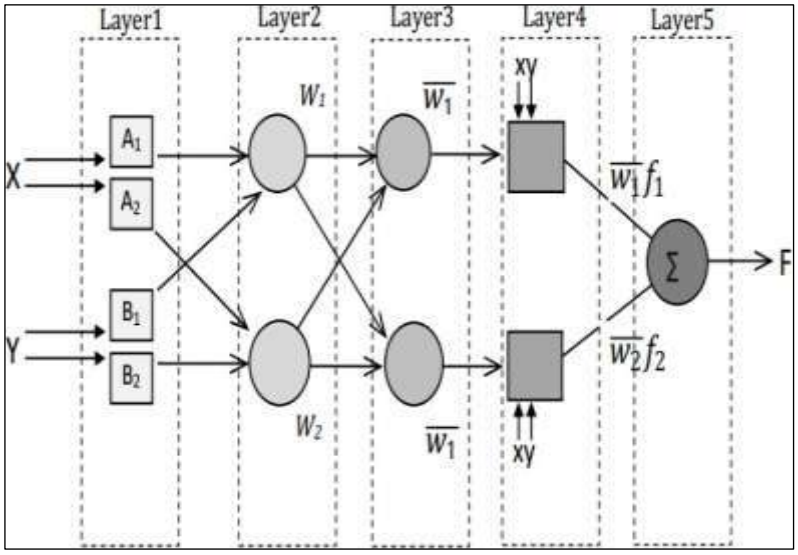


Figure 5: Rule based Adaptive Nuero Fuzzy Inference System.
Source: Authors, (2026).

From the fig.5 it is observe that the five-layer structure of a typical Sugeno-type ANFIS model.

Layer 1: Fuzzification Layer (Input Membership Functions)

This layer transforms the crisp input variables into fuzzy values by evaluating their membership degrees. Each neuron corresponds to a linguistic term such as Negative, Zero, or Positive. The output of this layer is:

$$O_{1,i} = \mu_{A_i}(x) \tag{12}$$

Where $\mu_{A_i}(x)$denotes the membership grade of input $x$to fuzzy set A_i. The parameters of this layer are called premise parameters, which are tuned during training.

Layer 2: Rule Layer (Firing Strength Computation)

Each neuron in this layer represents a fuzzy rule. The nodes multiply the incoming fuzzified signals to compute the firing strength of each rule. For a 2-input system:

$$O_{2,i} = w_i = \mu_{A_i}(x) \cdot \mu_{B_i}(y) \tag{13}$$

Thus, this layer models fuzzy logic's AND operation using the product operator.

Layer 3: Normalization Layer

This layer normalizes the firing strengths of all rules to ensure they sum to one. The normalized weight for rule $i$is calculated as:

$$O_{3,i} = \bar{w}_i = \frac{w_i}{\sum_{k=1}^{n}.w_k} \tag{14}$$

This normalization reflects the contribution level of each rule in the inference process.

Layer 4: Consequent Layer (Sugeno Rule Output)

Each neuron in this layer generates the rule's output based on a first-order Sugeno fuzzy model. The output of this layer is:

$$O_{4,i} = \bar{w}_i f_i = \bar{w}_i(p_i x + q_i y + r_i) \tag{15}$$

Here, p_i, q_i, and r_i are consequent parameters, trained together with premise parameters using a hybrid learning algorithm.

Layer 5: Output Layer (Defuzzification)

This layer aggregates the weighted outputs of all rules to obtain a single crisp system output. The output is computed as:

$$O_5 = \sum_{i=1}^{n}.\bar{w}_i f_i \tag{16}$$

Thus, Layer 5 performs defuzzification through a weighted average, delivering the final ANFIS prediction or control action. From section I to section IV are integrated based on the scope of the proposed control strategy of PMSM with EV system is tested in MATLAB Simulink Environment which described elaborately in the subsequent section in this paper.

V. TESTING OF SIMULINK BLOCK WITH PROPOSED METHOD

The Simulink block implemented by integrating of PMSM dc bus capacitor winding based with ESO and ANFIS which is shown in the fig.6 and it demonstrates the dual benefits of the winding-based capacitor discharge method and the ANFIS speed regulation strategy for PMSM-driven electric vehicles. The parameters of the PMSM Simulink model as shown in table.2. The Simulink block consists of the main functional subsystems required for the proposed robust dc-bus capacitor discharge control. The power supply section includes the fuel cell and battery connected through a bidirectional DC/DC converter. The power inverter drives the Permanent Magnet Synchronous Machine (PMSM) using space-vector pulse-width modulation (SVPWM).

The motor control strategy is implemented using an ANFIS-based speed controller along with d-q axis current controllers for accurate torque and current tracking. An Extended State Observer (ESO) is integrated to estimate disturbances and maintain dc-bus voltage stability during discharge events. Additionally, an electric braking interface and sensor feedback loops are included to ensure safety and smooth operational transitions between normal driving and discharge mode. From stable operation the EV battery fuel systems charging with an SOC condition. Once discharge the system automatically switches back to normal operation, where the fuel cell supplies power to the dc-bus and recharges the battery. This ensures a smooth transition and efficient recovery of the vehicle's energy.

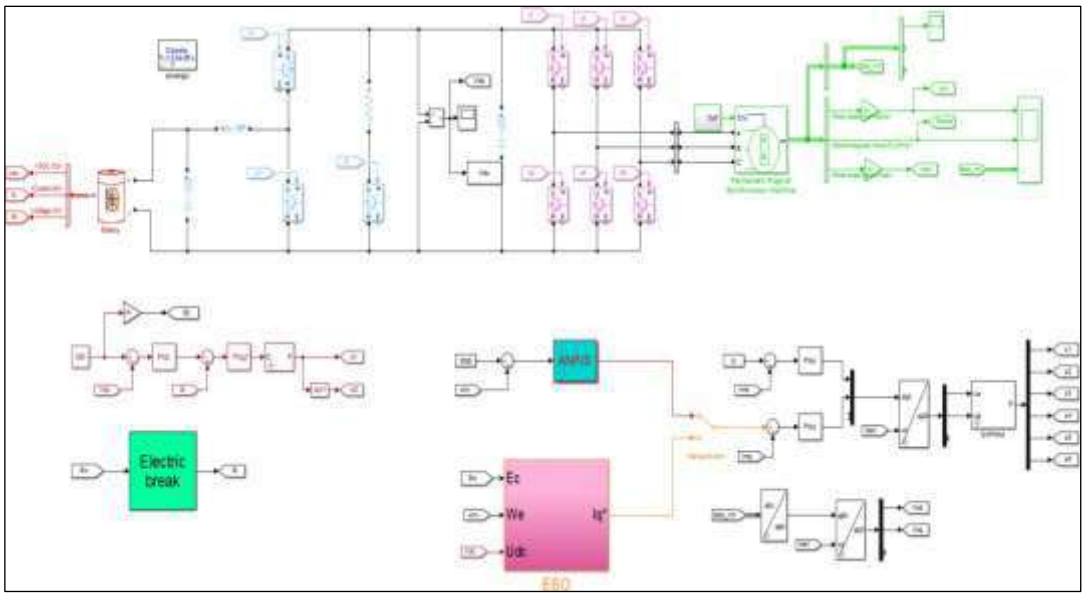


Figure 6: Simulation designing of proposed PMSM drive using ANFIS for EV.
Source: Authors, (2026).

Table:2 Parameters of PMSM.

Parameter	Symbol	Value
DC Bus Capacitor	C	420μ_F
DC Bus Voltage	V_{dc}	320V
Stator Resistance	R_s	0.2 Ω
d-axis Inductance	L_d	1.5 mH
q-axis Inductance	L_q	2.0 mH
Permanent Magnet Flux	ψ_f	0.12 Wb
Pole Pairs	p	4
Rotor Inertia	J	0.01 kgm²
Load Torque	T_{load}	10 - 50 Nm
Motor speed	ω_m	1000rpm

Source: Authors, (2026).

During the emergency scenario, the capacitor discharge process was initiated by injecting a large flux- enfeebling current through the stator windings, where the Extended State Observer (ESO) played a vital role in estimating system uncertainties such as copper losses, switching effects, and residual inductive energy. By compensating for these disturbances, the ESO ensured that the dc-bus voltage was held within a safe operating margin, preventing any rebound or unsafe levels of stored energy. However the Simulink model is tested and the corresponding test results are analyzed based on the speed regulation, torque, stator voltage, stator current as a case study 1. Steady state operation and 2. Dynamic state operation.

Case:I The steady state behavior of the PMSM drive system: ANFIS with EV

In this case, the steady-state performance of the ANFIS-controlled PMSM drive is evaluated at 1000 rpm under constant load that is 25 Nm. The controller accurately maintained the reference speed with negligible error and ensured stable dc-bus voltage (320V), battery voltage (130V), and current (25A,). The dq-axis currents remained steady, providing effective flux and torque control, while the electromagnetic torque showed 25Nm output with minimal ripple. The corresponding test output results speed characteristics 7(a), torque characteristics 7(b), DC bus voltage 7(c) and stator current characteristics 7(d) are shown in fig.7 under steady state operation.

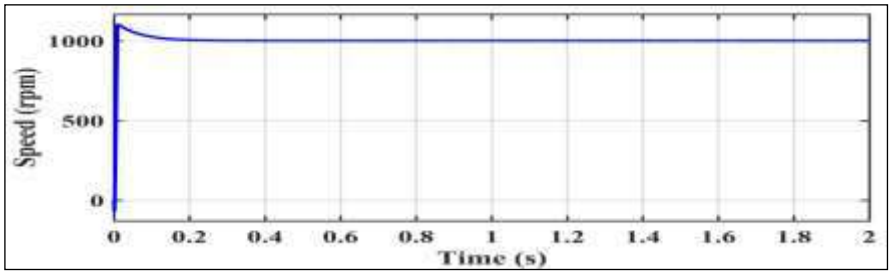


Figure 7(a): Speed(rpm) v/s Time(s).

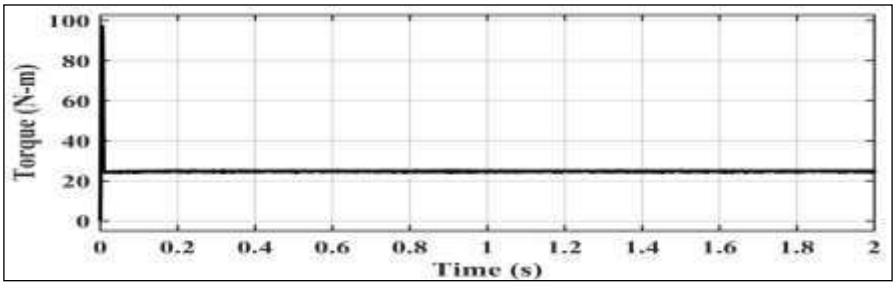


Figure 7(b): Torque(N-m) v/s Time(s).

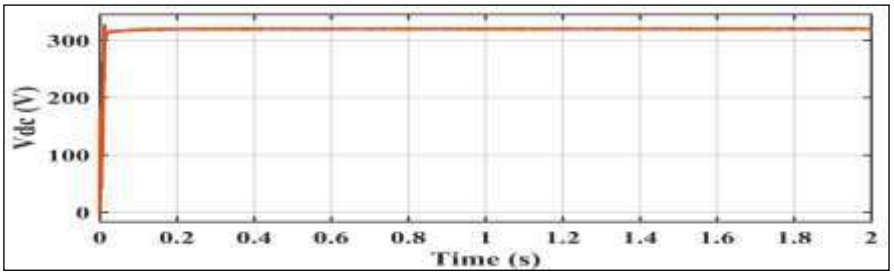


Figure 7(c): V_{dc} (v) v/s Time (s).

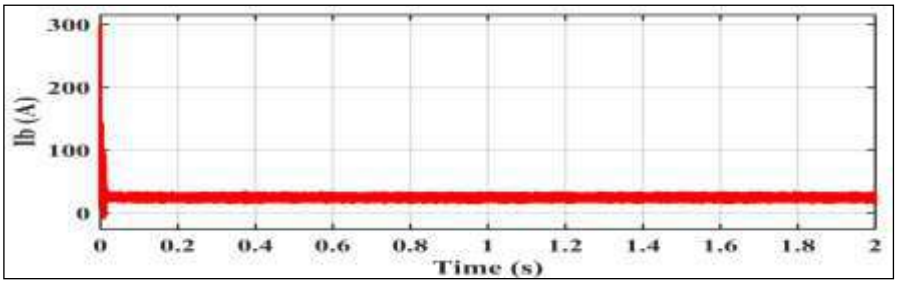


Figure 7(d): Ib (A) v/s Time (s).

Figure 7: Simulation results of the PMSM drive under steady-state operation.
Source: Authors, (2026).

Fig.7 shows that the Simulation test results of the PMSM drive under steady-state operation. Fig7(a), it is observed that PMSM with EV drive system initially the speed raises above the rated speed because of the rapid acceleration of torque and then the 0.2s speed is stabilizes and it reaches the rated speed of 1000 rpm proportionately maintain constant, validating its effectiveness for EV DC-bus capacitor discharge. Fig7(b) shows torque response reached peak value approximately 95Nm without damping oscillations because of the controller applies maximum acceleration torque to reach the reference speed rapidly, After the motor reaches the constant speed, therefore, the electromagnetic torque decreases and confirming stable performance of torque 25Nm during DC-bus discharge in EV. Fig 7(c) shows DC-bus voltage reached 340V because of the DC-bus capacitor draws a large inrush current.

After charging, the control system regulates the bus voltage, causing it to reduce and stabilize at the rated level and then maintained constant 320V, indicating stable regulation during capacitor discharge in EV. Fig 7(d) shows that the stator current initialy reached 300A peak value because of the motor needs a big initial torque to start moving, so the controller sends a large current for a short time and the fraction of seconds will falldown to low steady value and maintaining constant for confirming safe and controlled discharge in the EV PMSM drive. and then decreased 50A with high settling time and maintained constantly without any damping oscillations.

Case:II The dynamic behavior of the PMSM drive system: ANFIS for EV

The PMSM drive system under ANFIS control simulation model is tested with multiple speed transitions under a constant load that is 25Nm. The motor operated at different speeds 1000 rpm, 500 rpm, 300 rpm, 600 rpm, 800 rpm and 1000rpm. Throughout these variations, the ANFIS controller ensured fast speed tracking with negligible steady-state error and minimal overshoot. The dq-axis currents adapted smoothly, maintaining flux and torque control, while the electromagnetic torque showed minimal ripple. The dc-bus voltage (320V), battery voltage (130V), and current (25A,) remained stable, confirming effective power management.

The ANFIS-controlled PMSM drive demonstrated excellent dynamic performance and stability, making it well-suited for electric vehicle. The corresponding test output results speed characteristics 8(a), torque characteristics 8(b), DC bus voltage 8(c) and stator current characteristics 8(d) are shown in fig.8 under dynamic state operation.

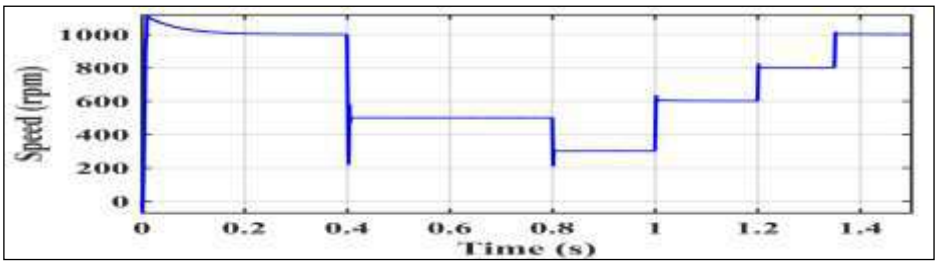


Figure 8(a): Speed(rpm) v/s Time(s).

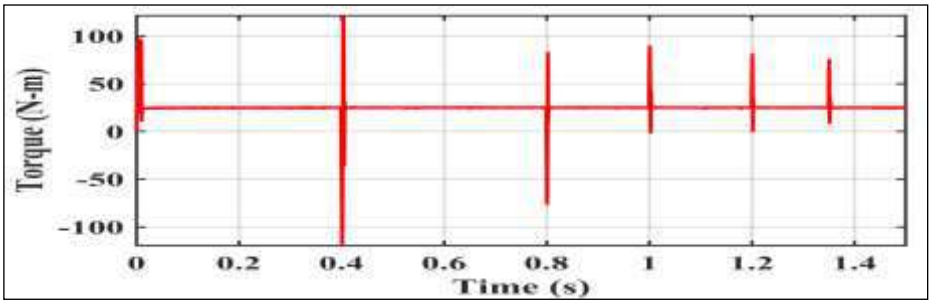


Figure 8(b): Torque (N-m) v/s Time(s).

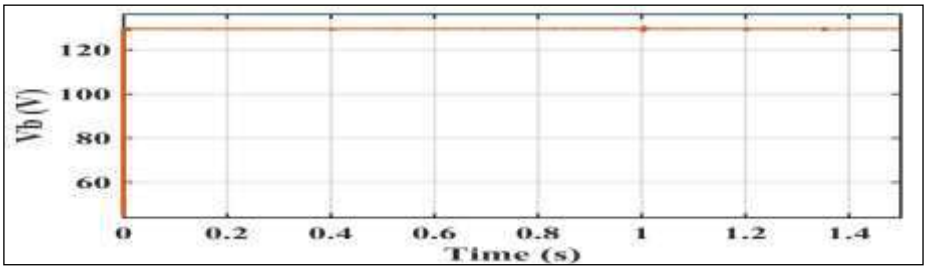


Figure 8(c): Vb (v) v/s Time(s).

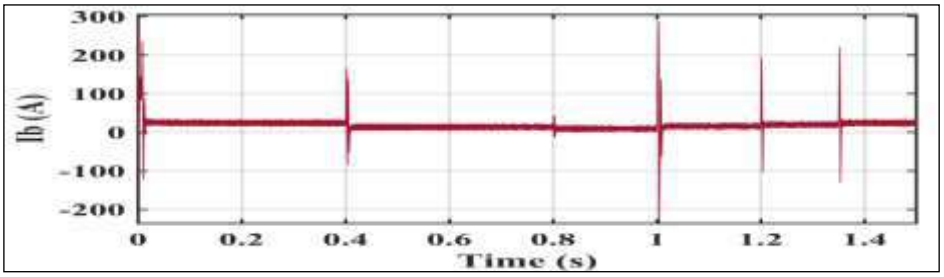


Figure 8(d): Ib(A) v/s Time(s).

Figure.8: Simulation Results under Dynamic Behavior of the PMSM Drive System.
Source: Authors, (2026).

Figure 8 shows that the Simulation test results of the PMSM drive under dynamic state operation. From the fig.8(a) observed that initially the speed shows peak value in a fraction of seconds and slightly settles the rated speed 1000 rpm because of the rapid acceleration of torque and it continuous and slowly step down to 500rpm at 0.4 s, after that it will continues and step down to 300rpm at 0.8s proportionately with 0.2 s time intervals the speed can be reaches to 1000rpm at 1.6s this process is analyzing until to reach the steady state. The motor tracks each reference level accurately, confirming strong dynamic response in EV drive conditions. Multiple speeds are used to evaluate controller performance under varying operating points.Fig. 8(b) shows torque variations in both positive and negative directions at different time intervals like 0.1s the load torque is 95Nm and then decreases to the statutory load limit 25Nm maintained with constant without damping oscillations at 0.4s, 0.8s and 1.4s respectively it indicating smooth transitions between motoring and regenerative modes with fast and controlled response.

Figure 8(c) shows that the DC-bus voltage stays around 125 V, confirming stable regulation during EV capacitor discharge. Fig 8(d) shows the range of current 20 A during normal operation. The speed reference changes, a sharp transient current spike at 0.4 s raised to nearly +180 A and then drops to –150 A as the motor responds to the speed reduction. The transient peaks at 0.8 s, 1.0 s, 1.2 s, and 1.4 s, where the current briefly reached values between 150-200 A, after each transient, the current quickly returns to its steady value around 10-20 A discharge current stays close to zero with different intervals only brief switching transients, demonstrating effective current regulation and stable capacitor discharge in the EV PMSM drive.

VI. ESO CONTROLLER FOR WINDING BASED SAFE DISCHARGE OF DC-BUS CAPACITOR

The simulation model of ESO controller as shown in fig.9. The Extended State Observer (ESO) controller is evaluated for the proposed winding-based safe discharge method in PMSM drives. During emergencies, the stator windings dissipate the dc-bus capacitor's stored energy after an initial flux-enfeebling d-axis current triggers rapid voltage reduction. The ESO accurately estimated and compensated for disturbances such as copper and switching losses, ensuring smooth, stable voltage decay without rebound. The simulation test results of the system as shown in fig 10(a) speed characteristics, fig 10(b) bus voltage characteristics, fig 10(c) torque respectively.

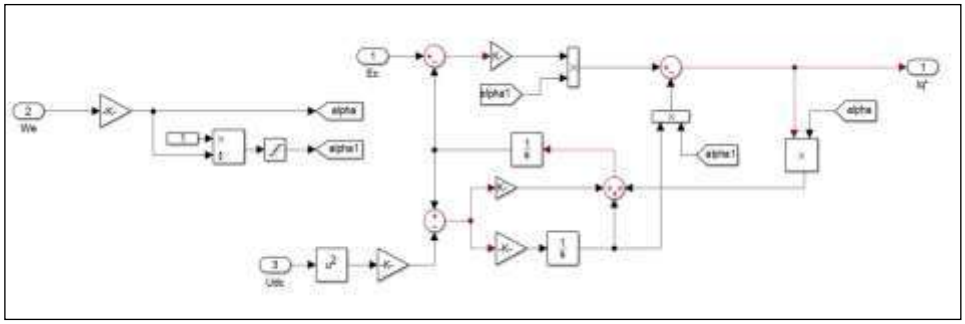


Figure 9: Simulation mode of ESO controller.
Source: Authors, (2026).

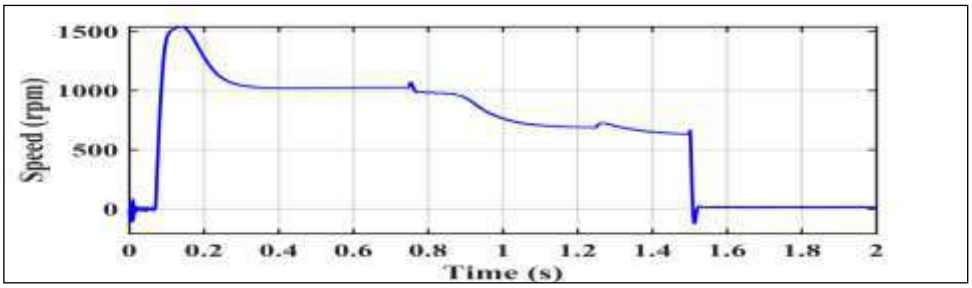


Figure 10(a): Speed(rpm) v/s Time(s).

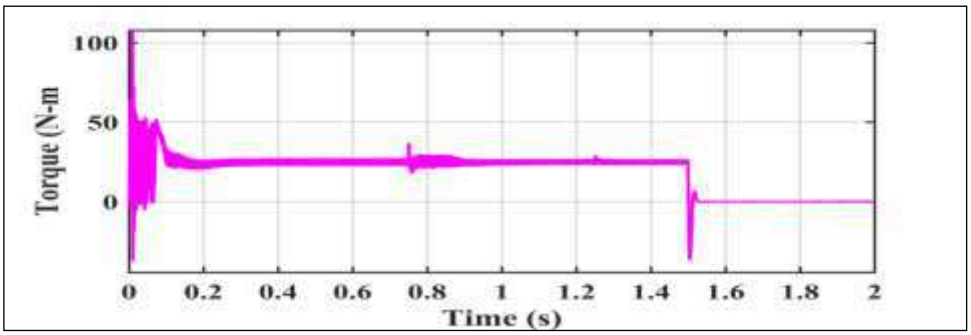


Figure 10(b) : Torque (N-m) v/s Time(s).

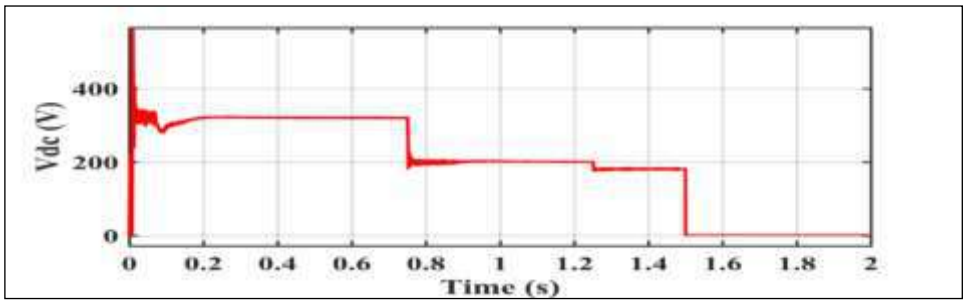


Figure 10(c): Vdc(v) v/s Time(s).

Figure 10: Simulation Results of ESO-Based Winding DC-Bus Capacitor for Safe Discharge.
Source: Authors, (2026).

From the simulation test results it is observed that fast discharge to safe voltage levels within limits compared to conventional methods, the ESO based approach achieved faster response, better disturbances rejection, and higher robustness confirming its effectiveness for safe PMSM operation in EV fault conditions. Fig.10 shows that the Simulation test results of the ESO controller for winding based safe discharge of dc bus capacitor. From the fig10(a) it is observed that PMSM speed during discharge the speed raises high peak around 1500 rpm due to the fast corrective action of the ESO, stabilizes near 1000 rpm, then gradually decreases in steps until it reaches zero, indicating controlled deceleration.

Figure 10(b) it is observed that the torque initially raised 100 Nm due to the startup transient and then settles around 40-45 Nm. A small disturbance appears near 0.8 s, 1.45 s the torque drops sharply to nearly 0 Nm, after it remains constant. Fig10(c) it is observed that the DC-bus voltage raised 390 V due to the initial transient, then stabilizes 300 V. At 0.8 s and 1.2 s it drops in steps to around 200 V and 170 V, and finally falls to 20 V at 1.4 s, indicating full capacitor discharge. Finally from the simulation test results are summarized compared to conventional methods with that of the proposed method which is tabulated in table.3.

Table:3 Comparison of test results.

Performance Criteria	PI Controller	Fuzzy Controller	ANN Controller	ANFIS Controller
Steady-State Speed (1000 rpm)	Acceptable regulation but small steady-state error present	Improved error reduction vs PI; dependent on rule accuracy	Good accuracy with reduced ripple but training-dependent	**Negligible error; precise tracking of 1000 rpm**
Dynamic Speed Transitions(1000 → 500 → 300 → 600 → 800 → 1000 rpm at 25 Nm load)	Slow response with overshoot and long settling	Faster than PI but moderate oscillations	Fast response; oscillations still exist	**Fastest response; minimal overshoot; shortest settling time**
Torque Ripple (25Nm)	High ripple	Moderate reduction depending on rules	Reduced ripple but still noticeable	**Smooth torque; minimal ripple**
dq-Axis Current Stability(25A)	Moderate stability	Good stability	Sensitive to parameter/training mismatch	**Excellent stability**
DC-Bus Voltage Regulation (320 V)	Poor stability during variations	Moderate performance	Good but with small fluctuations	**Stable and well-regulated at 320V**
Battery Operating Stability(130 V, 25 A)	Stable but slower dynamics	Good adaptation	Some fluctuations under disturbances	**Stable with strong disturbance rejection**

Source: Authors, (2026).

From the table.3 it is observed that the overall performance of the proposed control strategy for PMSM with EV system.

PI Controller: Shows acceptable steady-state response but suffers from slow transients and higher torque ripple. Stability reduces when dc-bus voltage drops during discharge.

Fuzzy Controller: Improves speed tracking and reduces ripple compared to PI. However, performance depends strongly on rule tuning and operating conditions.

ANN Controller: Provides faster dynamic response and better current regulation. Stability is affected by parameter variations outside the trained region, causing small fluctuations during discharge events.

ANFIS Controller: Delivers the best overall performance with fast settling, negligible speed error, smooth torque, and stable dq-axis currents maintains strong robustness against dc-bus capacitor discharge disturbances in EV operation.

VII CONCLUSIONS

The proposed control strategy discharge mechanism operates through two distinct stages to ensure rapid and controlled voltage reduction effectively addressed the limitations with active discharge circuits. The Simulation results confirmed that the two-stage discharge process, consisting of rapid flux-enfeebling current injection followed by ESO regulated voltage control, ensured a fast reduction of the dc-bus voltage to safe levels while preventing current and voltage overshoot. However the winding-based safe discharge method for the dc-bus capacitor in PMSM-driven electric vehicles, supported by ESO-based regulation and ANFIS-based speed control of electric vehicle.

The ESO demonstrated strong capability in estimating and compensating disturbances. However the integration of ESO based capacitor discharge with ANFIS based speed regulation provides a comprehensive solution that enhances both safety during emergencies and performance during normal driving operations. The proposed method shows strong potential for practical deployment in electric vehicle drive systems, ensuring passenger safety, reliable operation, and improved energy efficiency. Finally the proposed methodology ensures that safe voltage discharge during fault scenarios and enhanced the performance during standard operation, offering a reliable, intelligent and cost efficient solution for modern electric vehicle drive system.

VIII. AUTHOR'S CONTRIBUTION

Conceptualization: Chevuru Mahima, Ch. Chengaiah.
Methodology: Chevuru Mahima, Ch. Chengaiah.
Investigation: Chevuru Mahima, Ch. Chengaiah.
Discussion of results: Chevuru Mahima, Ch. Chengaiah.
Writing – Original Draft: Chevuru Mahima, Ch. Chengaiah.
Writing – Review and Editing: Chevuru Mahima, Ch. Chengaiah.
Resources: Chevuru Mahima, Ch. Chengaiah.
Supervision: Chevuru Mahima, Ch. Chengaiah.
Approval of the final text: Chevuru Mahima, Ch. Chengaiah.

IX.REFERENCES

[1] S. Sakunthala, R. Kiranmayi and P. N. Mandadi, "A Review on Speed Control of Permanent Magnet Synchronous Motor Drive Using Different Control Techniques," 2018 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), Chennai, 2018, pp. 97-102, doi: 10.1109/ICPECTS.2018.8521574.

[2] V. M. Bida, D. V. Samokhyalov and F. S. Al-Mahturi, "PMSM vector control techniques – A survey," 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), Moscow, 2018, pp. 577-581, doi: 10.1109/EIConRus.2018.8317164

[3] Y. Kung, Risfendra, Y. Lin and L. Huang, "FPGA-based sensorless controller for PMSM drives using sliding mode observer and phase locked loop," 2016 International Conference on Applied System Innovation (ICASI), Okinawa, 2016, pp. 1-4, doi:10.1109/ICASI.2016.7539798.

[4] S. Chen and H. Hoai, "Studying an Adaptive Fuzzy PID Controller for PMSM with FOC based on MATLAB Embedded Coder," 2019 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), YILAN, Taiwan, 2019, pp. 1-2, doi: 10.1109/ICCE TW46550.2019.8991743.

[5] M. B. B. Sharifian, T. Herizchi and K. G. Firouzjah, "Field oriented control of permanent magnet synchronous motor using predictive space vector modulation," 2009 IEEE Symposium on Industrial Electronics & Applications, Kuala Lumpur, 2009, pp. 574-579, doi: 10.1109/ISIEA.2009.5356385.

[6] [Online] https://in.element14.com/motor-control-permanent-magnet-sync-motor-pmsm- technology

[7] F. Korkmaz, İ. Topaloğlu, M. F. Çakir and R. Gürbüz, "Comparative performance evaluation of FOC and DTC controlled PMSM drives," 4th International Conference on Power Engineering, Energy and Electrical Drives, Istanbul, 2013, pp. 705-708, doi: 10.1109/PowerEng.2013.6635696.

[8] P. Sala-Perez, S. Galceran-Arellano and D. Montesinos-Miracle, "A sensorless stable V/f control method for a five-phase PMSM," 2013 15th European Conference on Power Electronics and Applications (EPE), Lille, 2013, pp. 1-10, doi: 10.1109/EPE.2013.6631943.

[9] K. Yang, X. Yang, H. Xie, Y. Liu, Y. Zhang and X. Wei, "Stable sensorless V/f and cosφ =1 control for permanent magnet synchronous motor drives," 2014 17th International Conference on Electrical Machines and Systems (ICEMS), Hangzhou, 2014, pp. 3564-3568, doi: 10.1109/ICEMS.2014.7014107.

[10] G. Xie, K. Lu, S. K. Dwivedi, J. R. Rosholm and F. Blaabjerg, "Minimum-Voltage Vector Injection Method for Sensorless Control of PMSM for Low-Speed Operations," in IEEE Transactions on Power Electronics, vol. 31, no. 2, pp. 1785-1794, Feb. 2016, doi: 10.1109/TPEL.2015.2426200.

[11] M. Konghirun and L. Xu, "A Fast Transient-Current Control Strategy in Sensorless Vector Controlled Permanent Magnet Synchronous Motor," in IEEE Transactions on Power Electronics, vol. 21, no. 5, pp. 1508-1512, Sept. 2006, doi: 10.1109/TPEL.2006.882419.

[12] S. Li, H. Won, X. Fu, M. Fairbank, D. C. Wunsch and E. Alonso, "Neural-Network Vector Controller for Permanent-Magnet Synchronous Motor Drives: Simulated and Hardware-Validated Results," in IEEE Transactions on Cybernetics, vol. 50, no. 7, pp. 3218-3230, July 2020, doi: 10.1109/TCYB.2019.2897653.
[13] S. Fartash Toloue, S. H. Kamali and M. Moallem, "Multivariable sliding-mode extremum seeking PI tuning for current control of a PMSM," in IET Electric Power Applications, vol. 14, no. 3, pp. 348-356, 3 2020, doi: 10.1049/iet-epa.2019.0033.

[14] In-Cheol Baik, Kyeong-Hwa Kim and Myung-Joong Youn, "Robust nonlinear speed control of PM synchronous motor using boundary layer integral sliding mode control technique," in IEEE Transactions on Control Systems Technology, vol. 8, no. 1, pp. 47-54, Jan. 2000, doi: 10.1109/87.817691. 89

[15] D. Liang, J. Li, R. Qu and W. Kong, "Adaptive Second-Order Sliding-Mode Observer for PMSM Sensorless Control Considering VSI Nonlinearity," in IEEE Transactions on Power Electronics, vol. 33, no. 10, pp. 8994-9004, Oct. 2018, doi: 10.1109/TPEL.2017.2783920.

[16] Mohamed, Y.A.I. Design and Implementation of a Robust Current-Control Scheme for a PMSM Vector Drive With a Simple Adaptive Disturbance Observer. IEEE Transactions on Industrial Electronics 2007, 54, 1981–1988.

[17] T. Türker, U. Buyukkeles and A. F. Bakan, "A Robust Predictive Current Controller for PMSM Drives," in IEEE Transactions on Industrial Electronics, vol. 63, no. 6, pp. 3906-3914, June 2016, doi: 10.1109/TIE.2016.2521338.

[18] S. Carpiuc and C. Lazar, "Fast Real-Time Constrained Predictive Current Control in Permanent Magnet Synchronous Machine-Based Automotive Traction Drives," in IEEE Transactions on Transportation Electrification, vol. 1, no. 4, pp. 326-335, Dec. 2015, doi: 10.1109/TTE.2015.2482223.
[19] W. Wang, Y. Fan, S. Chen and Q. Zhang, "Finite control set model predictive current control of a five-phase PMSM with virtual voltage vectors and adaptive control set," in CES Transactions on Electrical Machines and Systems, vol. 2, no. 1, pp. 136-141, March 2018, doi: 10.23919/TEMS.2018.8326460.

[20] G. Prior and M. Krstic, "Quantized-Input Control Lyapunov Approach for Permanent Magnet Synchronous Motor Drives," in IEEE Transactions on Control Systems Technology, vol. 21, no. 5, pp.

[21] C. Gong, Y. Hu, C. Gan, G. Chen and M. Alkahtani, "Modeling, Analysis, and Attenuation of Uncontrolled Generation for IPMSM Based Electric Vehicles in Emergency," in IEEE Trans. Ind. Electron., vol. 67, no. 6, pp. 4453-4462, June 2020.

[22] J. Liu, C. Gong, Z. Han and H. Yu, "IPMSM Model Predictive Control in Flux-Weakening Operation Using an Improved Algorithm," in IEEE Trans. Ind. Electron., vol. 65, no. 12, pp. 9378 9387, Dec. 2018.

[23] Y. Zhang, J. Jin and L. Huang, "Model-Free Predictive Current Control of PMSM Drives Based on Extended State Observer Using Ultralocal Model," in IEEE Trans. Ind. Electron., vol. 68, no. 2, pp. 993-1003, Feb. 2021.

[24] M. Fliess and C. Join, "Model-free control," Int. J. Control, vol. 86, no. 12, pp. 2228–2252, 2013. [25] C. Lin, T. Liu, J. Yu, L. Fu and C. Hsiao, "Model-Free Predictive Current Control for Interior Permanent-Magnet Synchronous Motor Drives Based on Current Difference Detection Technique," in IEEE Trans. Ind. Electron., vol. 61, no. 2, pp. 667-681, Feb. 2014.

[25] C. Lin, T. Liu, J. Yu, L. Fu and C. Hsiao, "Model-Free Predictive Current Control for Interior Permanent-Magnet Synchronous Motor Drives Based on Current Difference Detection Technique," in IEEE Trans. Ind. Electron., vol. 61, no. 2, pp. 667-681, Feb. 2014.

[26] S. Zhou, J. Liu, L. Zhou and Y. Zhang, DQ Current Control of Voltage Source Converters with a Decoupling Method Based on Preprocessed Reference Current Feed-forward, in IEEE Trans. Power Electron., vol. 32, no. 11, pp. 8904-8921, Nov. 2017.

[27] X. Zhang and J. Yang, "A Robust Flywheel Energy Storage System Discharge Strategy for Wide Speed Range Operation," in IEEE Trans. Ind. Electron., vol. 64, no. 10, pp. 7862-7873, Oct. 2017.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 718-727. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3040

RESEARCH ARTICLE OPEN ACCESS

PERFORMANCE EVALUATION OF JSON LIBRARIES IN KOTLIN SERVER ENVIRONMENTS: JACKSON, GSON, AND KOTLINX SERIALIZATION

Achmad Arwan*[1], Julia Nur Fajrina[2] and Eriq Muhammad Adam Jonemaro[3]

[1, 2, 3]Department of Informatic Engineering, Faculty of Computer Science, Brawijaya University, Malang, Indonesia.

[1]https://orcid.org/0000-0002-4146-363X, [2]https://orcid.org/0009-0008-4817-3599, [3]https://orcid.org/0000-0001-8315-3210

Email: *arwan@ub.ac.id, juliarina@student.ub.ac.id, eriq.adams@ub.ac.id

ARTICLE INFO

Article History
Received: December 2, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Kotlin,
JSON libraries,
Jackson,
Server-side performance,
JSON processing.

ABSTRACT

The widespread adoption of Kotlin for server-side development has emphasized the importance of efficient JSON processing. This study aimed to empirically evaluate and compare the performance of three JSON libraries—Jackson, Gson, and KotlinX Serialization—in Kotlin-based server environments. Using structured experiments, the research measured execution time, memory utilization, and CPU consumption during parsing, serialization, and deserialization tasks involving nested JSON data with incremental complexity. Statistical analyses confirmed significant performance differences among libraries. KotlinX Serialization consistently outperformed Jackson and Gson across execution speed, memory efficiency, and CPU utilization, primarily due to its compile-time code generation approach. Jackson showed superior memory management and efficient serialization for smaller datasets, while Gson exhibited lower overall performance. These results offer critical guidance for developers selecting appropriate JSON libraries, highlighting KotlinX Serialization's suitability for high-performance applications. The findings significantly enhance the current understanding of JSON processing efficiency and suggest avenues for further investigation into additional libraries and real-world application contexts.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The rapid evolution of software technology has significantly elevated the importance of server-side development, as backend systems have become crucial in ensuring scalability, performance, and robustness of modern applications. Kotlin, a statically-typed programming language introduced by JetBrains, has steadily gained popularity as an efficient alternative to Java, particularly for server-side programming. Kotlin's concise syntax, enhanced type safety, interoperability with Java, and modern programming paradigms such as coroutines have substantially increased its adoption among developers [1]. This growth has been notably supported by frameworks explicitly tailored for Kotlin, such as Ktor, and by mainstream Java frameworks like Spring Boot, which have adopted official support for Kotlin [2]. According to the JetBrains Developer Ecosystem Report (2023), approximately 52% of Kotlin developers use Kotlin for backend development, signifying an upward trend in its acceptance within the industry [2].

Leading global enterprises including Amazon, Atlassian, Adobe, and Mercedes-Benz.io have embraced Kotlin within their backend infrastructure, particularly for microservices and enterprise-level applications, further validating its suitability for high-performance environments [3]. A crucial aspect of server-side development involves effective data communication, where JSON (JavaScript Object Notation) has become the standard format due to its lightweight nature, ease of parsing, human readability, and compatibility across diverse platforms[4],[5]. JSON's semi-structured nature makes it adaptable to various application scenarios but also requires efficient processing mechanisms, including parsing, serialization, and deserialization, to transform JSON data into usable native objects within server applications. These processing tasks significantly impact the overall performance of a server, particularly in contexts involving high volumes of concurrent data exchanges [6]. Efficient JSON processing is thus imperative, highlighting the importance of selecting suitable JSON libraries capable of optimizing performance and resource utilization in Kotlin-based server environments.

Despite the availability of numerous JSON processing libraries, empirical comparisons that specifically focus on Kotlin server environments remain relatively scarce. The core research problem arises from the necessity to evaluate performance discrepancies among available JSON libraries under realistic workloads typical of Kotlin server applications. Different libraries approach JSON processing through varying computational strategies. Prominent among these are Jackson, Gson, and KotlinX Serialization, each employing distinct methodologies that directly influence their runtime efficiency. Jackson and Gson predominantly rely on reflection, dynamically inspecting objects at runtime to facilitate JSON transformations[7],[8]. Conversely, KotlinX Serialization employs a fundamentally different strategy, generating serialization logic at compile-time through compiler plugins, thus reducing runtime overhead [9].

Addressing this central problem requires an approach encompassing empirical evaluation, systematic experimentation, and rigorous statistical analysis. General solutions often involve benchmarking libraries against multiple performance metrics, including execution time, memory consumption, and CPU usage, under various data complexity scenarios. However, such evaluations must also explicitly consider statistical significance to ensure robust, scientifically valid conclusions. Existing literature typically lacks this methodological rigor or generalizes findings from non-Kotlin environments, reducing their relevance to the Kotlin ecosystem. Previous research provides insights into JSON processing performance, although these studies differ substantially in terms of contexts, methodologies, and goals.

Maeda examined the efficiency of various serialization formats, including XML, JSON, and binary formats, within Java-based applications, focusing on comparative metrics such as serialization speed and output size[10]. Vanura and Kriz further evaluated the performance of JSON, MessagePack, and Protocol Buffers across languages like Java, PHP, and JavaScript, emphasizing both speed and data compactness[11]. Similarly, Dhalla conducted performance comparisons among native JSON parsers across multiple languages—Java, Python, .NET Core, JavaScript, and PHP—with emphasis on parsing efficiency, memory usage, and CPU consumption [5]. While valuable, these studies either omit Kotlin-specific implementations or neglect rigorous statistical testing, presenting limitations that hinder direct applicability to Kotlin-based server contexts.

Moreover, in studies specifically addressing Kotlin, KotlinX Serialization emerges as a promising solution explicitly designed to address Kotlin's inherent language features and runtime characteristics. Unlike Jackson and Gson, KotlinX Serialization utilizes compiler-generated code, providing native support for Kotlin's language-specific attributes such as data classes, default parameters, and null-safety guarantees, significantly reducing runtime reflection overhead and improving serialization performance[12]. This inherent compatibility and performance optimization through compile-time code generation positions KotlinX Serialization as an advantageous choice, particularly in scenarios demanding rapid processing and minimal resource usage. However, while KotlinX Serialization theoretically offers these advantages, empirical evidence supporting these claims in realistic server-side conditions remains insufficiently explored.

Previous research has generally tested JSON processing performance within generic or language-neutral environments, often neglecting Kotlin's specific runtime interactions and JVM integration nuances. Furthermore, the absence of comprehensive experimental scenarios involving complex, nested JSON structures, which closely resemble practical server-side data exchanges, limits the existing evidence's real-world applicability. Consequently, a clearly defined research gap emerges, characterized by inadequate empirical benchmarking of JSON libraries within the Kotlin server ecosystem, specifically regarding rigorous statistical testing and practical experimental scenarios involving nested and complex data structures. The existing literature demonstrates significant variations in performance among libraries across different environments and programming contexts, yet lacks specificity and methodological rigor for Kotlin-based implementations.

To address this gap, this study aims to empirically evaluate and compare the performance of three widely-used JSON libraries—Jackson, Gson, and KotlinX Serialization—in realistic server-side scenarios explicitly developed for Kotlin. The primary objective is to systematically assess and quantify each library's efficiency in parsing, serialization, and deserialization tasks involving nested JSON data of varying complexities. The experimental design incorporates multiple performance metrics, namely execution time, memory usage, and CPU consumption, measured under strictly controlled, reproducible conditions. Furthermore, the research methodology integrates comprehensive statistical analysis to validate the significance of observed performance differences, enhancing the robustness and scientific validity of the findings.

The novelty of this study resides in its dedicated Kotlin-specific focus, addressing the need for evidence-based insights applicable directly to Kotlin server-side development. By employing rigorous statistical methodologies and realistic, structured experimental scenarios, the research seeks to provide definitive, actionable insights to support informed decision-making among developers selecting JSON processing libraries. The scope of the study is explicitly confined to server-side JSON processing within Kotlin-based environments, excluding mobile and front-end applications, thereby offering targeted relevance and practical utility for Kotlin backend developers and researchers.

II. THEORETICAL REFERENCE

II.1 SERVER-SIDE DEVELOPMENT AND THE ROLE OF KOTLIN

Server-side development has become fundamental to modern software architecture, providing the backbone for scalability, performance optimization, and system robustness. Kotlin, a statically-typed programming language developed by JetBrains, has emerged as a compelling alternative to Java for backend systems. According to the JetBrains Developer Ecosystem Report (2023), approximately 66% of Kotlin developers utilize the language for Android and server-side applications, demonstrating substantial adoption in enterprise environments [13], [14].

Kotlin's design philosophy emphasizes concise syntax, enhanced type safety, and seamless interoperability with existing Java ecosystems. The language incorporates modern programming paradigms, particularly coroutines for asynchronous programming, which enable developers to write non-blocking code in a sequential style without the complexity of traditional callback mechanisms. Spring Boot, one of the most widely adopted backend frameworks, has provided official Kotlin support since 2017, offering first-class coroutine support, Kotlin-specific DSLs, and enhanced null-safety features.

Similarly, Ktor, a lightweight asynchronous framework built by JetBrains specifically for Kotlin, leverages coroutines for high scalability and provides intuitive Domain-Specific Language (DSL) capabilities for server-side development. Major global enterprises including Amazon, Atlassian, Adobe, and Mercedes-Benz have integrated Kotlin into their backend infrastructure, particularly for microservices and enterprise-level applications, validating its suitability for production environments requiring high performance and reliability. Type safety represents a fundamental property of programming languages that ensures variables access only authorized memory locations in well-defined, permissible ways, thereby preventing invalid operations on objects.

Statically-typed languages like Kotlin perform type checking at compile-time, enabling compilers to optimize code execution and memory usage while catching type errors before runtime. Kotlin's type system incorporates null-safety as a first-class language feature, distinguishing between nullable and non-nullable types at the type-system level. This design choice eliminates entire categories of runtime errors common in Java, particularly NullPointerException occurrences. For JSON serialization libraries, native support for Kotlin's type system features—as provided by KotlinX Serialization—theoretically reduces impedance mismatch between language semantics and serialization logic[15].

II.2 JSON AS THE STANDARD FOR DATA EXCHANGE

JavaScript Object Notation (JSON) has established itself as the de facto standard for data interchange in server-side development due to its lightweight structure, human readability, and cross-platform compatibility. JSON's semi-structured nature provides flexibility for representing complex data hierarchies while maintaining simplicity in parsing and generation. Data serialization, defined as the process of converting data objects into byte streams or character sequences for storage or transmission, forms a critical component of server-side operations. The reverse process, deserialization, reconstructs the original data structure from the serialized format.

In distributed systems and high-performance computing environments, efficient serialization directly impacts system throughput, latency, and resource utilization[16], JSON serialization involves encoding complex data structures while preserving information such as type, order, and relationships between nested elements. The performance characteristics of JSON processing—including parsing speed, memory consumption, and CPU overhead—significantly affect overall server performance, particularly in scenarios involving high-volume concurrent data exchanges typical of modern backend applications.

II.3 PERFORMANCE METRICS IN SOFTWARE SYSTEMS

Performance evaluation of software systems relies on standardized metrics that quantify computational efficiency and resource utilization. The three primary performance dimensions assessed in this study are Execution Time, Memory Usage and CPU Consumption. Execution Time (CPU Time) represents the total duration a CPU spends computing a specific task, excluding time spent on I/O operations or executing other processes. CPU execution time can be calculated as the product of CPU clock cycles and clock cycle time, or equivalently, as the ratio of instruction count and clock rate. This metric directly measures computational efficiency and is inversely proportional to system performance.

Memory Usage quantifies the amount of RAM consumed during program execution, typically measured in bytes, kilobytes, or megabytes. Memory utilization patterns significantly impact system scalability, particularly in server environments handling concurrent requests. Excessive memory consumption can trigger garbage collection overhead in JVM-based systems, indirectly affecting execution time. CPU Consumption is the Measures the percentage of CPU resources utilized during execution, providing insight into processor workload and potential bottlenecks. High CPU utilization may indicate computationally intensive operations, while low utilization might suggest I/O-bound or inefficiently parallelized code.

II.4 STATISTICAL SIGNIFICANCE TESTING IN PERFORMANCE EVALUATION

Statistical significance testing provides the methodological framework for determining whether observed performance differences between systems reflect genuine effects rather than random variation. The Neyman-Pearson hypothesis testing paradigm, widely adopted in empirical software engineering, involves formulating a null hypothesis (typically asserting no difference between conditions) and an alternative hypothesis, then calculating the probability (p-value) of obtaining the observed data assuming the null hypothesis is true [17]. A critical consideration in performance benchmarking is the relationship between sample size and statistical power. As sample size increases, the probability of detecting true effects (statistical power) increases while the risk of Type II errors (false negatives) decreases.

However, with very large samples, even trivial effect sizes may achieve statistical significance, necessitating consideration of practical significance in addition to statistical significance [17].The choice of significance threshold (α level, commonly set at 0.05) represents a balance between Type I errors (false positives) and Type II errors (false negatives). Some researchers argue that exploratory studies with small sample sizes may warrant more liberal thresholds (e.g., $\alpha = 0.20$), while confirmatory studies with large samples should employ more conservative thresholds. Beyond p-values, effect size measures quantify the magnitude of observed differences, providing context for practical importance independent of sample size[17].

II.5 PRIOR RESEARCH ON SERIALIZATION PERFORMANCE

Previous empirical studies have examined serialization format performance across various programming language contexts, though comprehensive Kotlin-specific evaluations remain limited. Maeda (2012) compared object serialization libraries across XML, JSON, and binary formats in Java-based applications, focusing on serialization speed and output size as primary metrics. Vanura and Kriz evaluated JSON, MessagePack, and Protocol Buffers across Java, PHP, and JavaScript, emphasizing speed and data compactness but omitting Kotlin implementations and rigorous statistical validation [18]. Conducted performance analysis of native JSON parsers across five programming languages—Java, Python, PHP, .NET Core, and JavaScript—measuring parsing speed, memory usage, and CPU consumption for JSON documents with varying nesting depth and key-value pair counts. Their findings demonstrated substantial performance variation across language environments, though Kotlin was not included in the comparative analysis [19].

Studies specifically examining MessagePack, Protocol Buffers, and other binary serialization formats have consistently found trade-offs between serialization speed, deserialization speed, and payload size. Protocol Buffers typically achieves faster serialization and smaller payloads compared to JSON, while MessagePack offers balanced performance across multiple dimensions. However, these studies generally employ Java-centric implementations without considering Kotlin-specific libraries or language features. Research addressing Kotlin serialization specifically highlights KotlinX Serialization's compile-time code generation approach as theoretically advantageous for performance. However, empirical validation under realistic server-side workloads with rigorous statistical analysis and comprehensive performance metrics remains underrepresented in the literature. This gap motivates the present study's focus on systematic, statistically validated comparison of Jackson, Gson, and KotlinX Serialization within Kotlin server environments [20].

III. MATERIALS AND METHODS

Write This study adopts a quantitative, empirical research approach to systematically evaluate and compare the performance of JSON libraries—Jackson, Gson, and KotlinX Serialization—within a Kotlin-based server environment. A between-subject experimental design was utilized, wherein each JSON library was tested independently against identical JSON data sets. This experimental approach was chosen to ensure unbiased performance metrics specific to each library, allowing precise analysis of differences in execution time, memory utilization, and CPU consumption under controlled conditions. The research diagram shown on Figure 1.

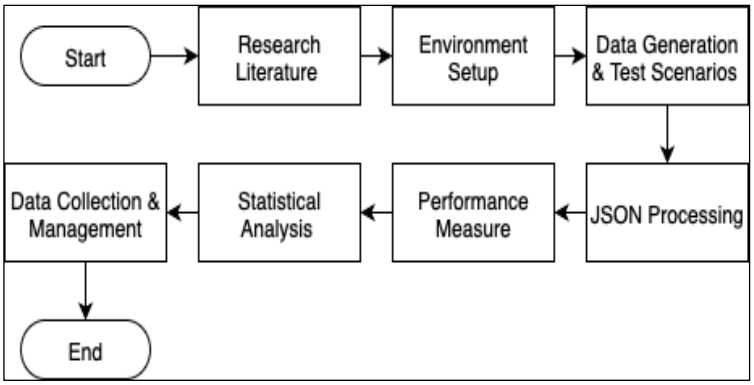


Figure 1: Research Methodology.
Source: Authors, (2026).

III.1 EXPERIMENTAL ENVIRONMENT

The experimental setup consisted of a server developed using the Kotlin programming language, facilitated by the Ktor framework, to ensure relevance to real-world server implementations. Docker containers were employed to guarantee consistency, reproducibility, and isolation of experimental environments across multiple test runs. The hardware specifications used in this research included an HP Pavilion 14 laptop with an AMD Ryzen 5 5500U processor, 16 GB RAM, and Windows 11 64-bit operating system. Precise measurement of execution time was obtained using the System.nanoTime() method. Memory utilization was assessed by comparing heap memory usage before and after each process through the Java MemoryMXBean API, and CPU consumption was monitored using OperatingSystemMXBean.

III.2 DATA GENERATION AND TEST SCENARIOS

Test data were generated using an adapted version of an algorithm developed by [5] to create structured, nested JSON files suitable for benchmarking JSON libraries. The data comprised ten distinct JSON files, each characterized by a fixed nesting depth of 10 levels, with increasing complexity achieved by incrementing the number of key-value pairs at each nesting level from 10 to 100, in steps of 10. Each file thus represented progressively larger and more complex data structures, reflective of realistic server-side processing scenarios. At the deepest nesting level, a structured JSON array containing additional nested objects extended the complexity further to an effective depth of 20 levels, emulating real-world data complexity. The structure data shown on Figure 2. The file size and the number of key shown on Table 1.

III.3 JSON PROCESSING PROCEDURES

The evaluation encompassed three primary JSON data processing tasks: parsing, serialization, and deserialization. Parsing refers explicitly to the transformation of raw JSON strings into intermediary data structures (JsonElement), without type-specific conversions. In contrast, deserialization converts JSON strings directly into typed Kotlin objects, while serialization transforms Kotlin objects back into JSON strings. These definitions align with distinctions drawn by [21] in the KotlinX Serialization documentation, ensuring clarity and consistency across all processing tasks.



Figure 2: JSON Data Structure.
Source: Authors, (2026).

Table 1: Key and file Size.

No	*Key Number*	*File Size* (KB)
1	10	58
2	20	117
3	30	176
4	40	236
5	50	295
6	60	355
7	70	414
8	80	474
9	90	533
10	100	592

Source: Authors, (2026).

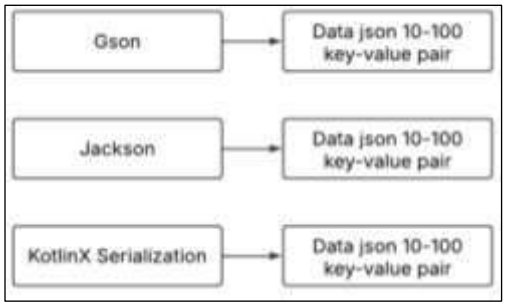


Figure 3: Library and Number Experiments.
Source: Authors, (2026).

Each processing task was performed in two stages to enhance accuracy and reliability: an initial warm-up phase followed by a main measurement phase. The warm-up phase consisted of 500 initial iterations intended to stabilize the JVM environment, mitigating the effects of Just-In-Time (JIT) compilation and runtime optimizations. Subsequently, the main measurement phase involved 5,000 experimental runs, each comprising 50 repetitions of parsing, serialization, or deserialization per test file. Performance metrics were averaged across runs to yield stable and representative results.

III.4 PERFORMANCE METRICS

Three key performance indicators were systematically measured: execution time, memory usage, and CPU consumption. Execution time, recorded in milliseconds (ms), was captured precisely to evaluate processing speed. Memory usage, assessed in megabytes (MB), provided insights into the libraries' efficiency in resource utilization. CPU consumption, expressed as a percentage, indicated computational load under identical processing tasks, offering a comprehensive view of resource efficiency. All metrics were recorded consistently across parsing, serialization, and deserialization processes to enable direct comparative analysis.

III.5 STATISTICAL ANALYSIS

Rigorous statistical analyses were conducted to ensure the reliability and validity of the performance comparisons. Initially, data were tested for normality using the Kolmogorov-Smirnov test to establish appropriate subsequent analytical methods. Due to the non-normal distribution of performance metrics, non-parametric statistical methods were selected for further analysis. Specifically, the Kruskal-Wallis test, a non-parametric alternative to ANOVA, was employed to detect statistically significant differences among library performances across all experimental conditions. A significance level (α) of 0.05 was chosen to determine the statistical significance of observed differences. Following the identification of significant differences via the Kruskal-Wallis test, Dunn's post hoc test was used for pairwise comparisons to pinpoint precisely where significant differences existed between individual library pairs[22].

III.6 DATA COLLECTION AND MANAGEMENT

Performance data obtained from experimental runs were systematically logged and stored in CSV format to facilitate structured data management and subsequent statistical analysis. Each CSV entry contained clearly labelled and timestamped metrics for execution time, memory utilization, and CPU consumption, ensuring traceability and reproducibility of data collection processes. Data integrity and consistency were verified through regular validation checks performed throughout the experimental period.

III.7 LIMITATIONS AND SCOPE

The methodological approach of this study was deliberately limited to evaluating three widely-adopted JSON libraries within a Kotlin-based server context. This decision was guided by the prevalence and relevance of these libraries within Kotlin's developer ecosystem. The scope excluded libraries not explicitly designed or widely used for Kotlin environments and did not extend to front-end or mobile applications. Furthermore, testing was confined to controlled laboratory conditions, and results may vary in different hardware or operating environments.

Future research could explore broader comparisons involving additional libraries, other programming environments, or real-world deployment scenarios. Overall, the structured and systematic approach adopted in this study, encompassing rigorous experimental procedures and comprehensive statistical analysis, was designed to yield valid, reliable, and practically relevant insights into the comparative performance of Jackson, Gson, and KotlinX Serialization libraries within Kotlin-based server environments. This methodology addresses a clear research gap, providing robust empirical evidence to guide informed decision-making among software developers and system architects.

IV. RESULTS AND DISCUSSIONS

IV.1 OVERVIEW OF EXPERIMENTAL RESULTS

This section presents the results obtained from rigorous empirical testing of three JSON libraries—Jackson, Gson, and KotlinX Serialization—within a Kotlin-based server environment. The performance metrics examined include execution time, memory usage, and CPU consumption, measured consistently across parsing, serialization, and deserialization processes. Each process was evaluated using structured nested JSON datasets, progressively increasing in complexity by incrementing key-value pairs from 10 to 100 per nesting level. Detailed statistical analyses were conducted to ensure the reliability and robustness of the findings.

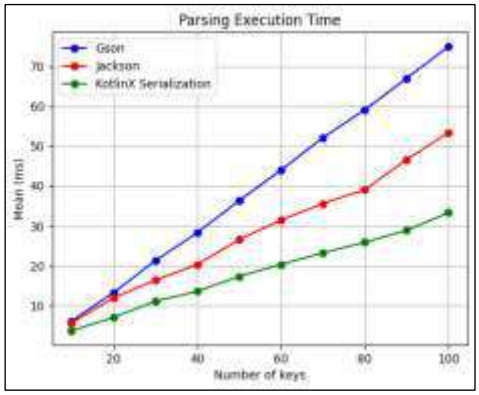


Figure 4: Parsing Performance.
Source: Authors, (2026).

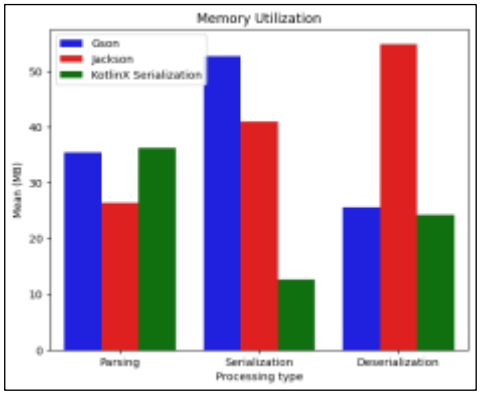


Figure 5: Memory Utilization.
Source: Authors, (2026).

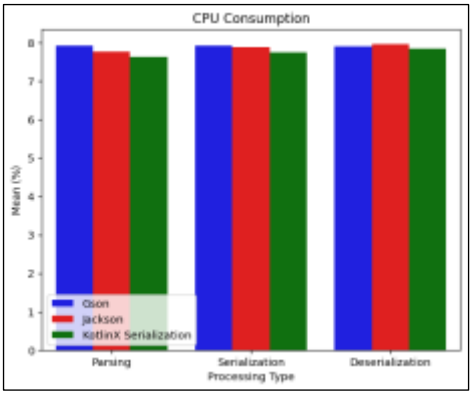


Figure 6: CPU Consumption.
Source: Authors, (2026).

IV.1.1 Parsing Performance

The parsing process involved transforming raw JSON strings into intermediate JSON data structures (JsonElement). The results of performance evaluation shown on Figure 4. Performance evaluation revealed distinct differences among the libraries tested. Specifically, KotlinX Serialization consistently exhibited the fastest execution times across all file complexities. For instance, with the smallest file (10 key-value pairs), KotlinX Serialization achieved an average execution time of 3.63 milliseconds, significantly outperforming Jackson (5.57 milliseconds) and Gson (6.03 milliseconds). As file complexity increased, performance advantages for KotlinX Serialization became increasingly pronounced, with a clear trend observed across all complexity levels. Memory utilization during parsing, serialize and deserialize depicted on Figure 5. Memory utilization during parsing showed a different pattern. Jackson demonstrated superior efficiency, consistently consuming less memory compared to both Gson and KotlinX Serialization.

For example, Jackson's average memory usage was approximately 26.35 megabytes, markedly lower than Gson's 35.46 megabytes and KotlinX Serialization's 36.27 megabytes across all tests. This efficiency advantage for Jackson remained consistent irrespective of file size or complexity. CPU consumption during parsing, serialize and deserialize depicted on Figure 6. CPU consumption data during parsing further emphasized KotlinX Serialization's performance advantage, albeit with narrower margins. KotlinX Serialization recorded the lowest CPU usage at an average of 7.01%, marginally better than Jackson (7.49%) and Gson (7.93%). This lower CPU consumption by KotlinX Serialization suggests better optimization in computational efficiency during parsing processes.

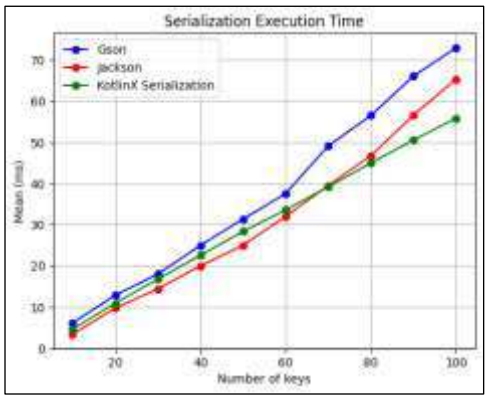


Figure 7: Serialization Performance.
Source: Authors, (2026).

IV.1.2 Serialization Performance

Serialization performance, involving the transformation of native Kotlin objects into JSON strings, presented contrasting results compared to parsing. The results of Serialization performance shown on Figure 7. Jackson emerged as the most efficient library regarding execution time for smaller and medium-sized files (10 to 60 key-value pairs). For instance, at the smallest file size (10 keys), Jackson completed serialization in an average of 3.36 milliseconds, significantly quicker than KotlinX Serialization's 4.56 milliseconds and Gson's 6.08 milliseconds. However, as file complexity grew (70 to 100 key-value pairs), KotlinX Serialization progressively outperformed Jackson and Gson, clearly demonstrating superior scalability for larger data sets.

Memory usage during serialization illustrated pronounced advantages for KotlinX Serialization across all file sizes (Figure 5). KotlinX Serialization consistently used significantly less memory, averaging 12.59 megabytes, compared to Jackson's 40.87 megabytes and Gson's 52.64 megabytes. This substantial efficiency in memory utilization highlights KotlinX Serialization's suitability for resource-constrained server environments. CPU consumption data indicated minimal variations between libraries, with KotlinX Serialization marginally outperforming others (depicted on Figure 6). It exhibited average CPU usage of 7.38%, slightly lower than Jackson (7.43%) and Gson (7.69%), signifying a minor yet consistent advantage in computational efficiency during serialization processes.

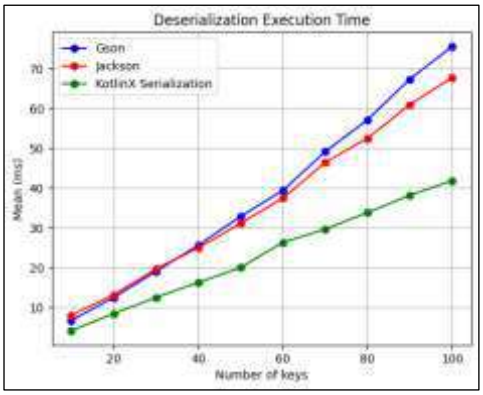


Figure 8: Deserialization Performance.
Source: Authors, (2026).

IV.1.3 Deserialization Performance

The deserialization process, converting JSON strings directly into typed Kotlin objects, underscored KotlinX Serialization's comprehensive performance superiority. The results of deserialization performance shown on Figure 8. Execution time results consistently favoured KotlinX Serialization across all file complexities, from simple to highly complex JSON structures. For instance, at the smallest complexity level (10 keys), KotlinX Serialization deserialized data in an average of 4.11 milliseconds, significantly faster than Gson (6.68 milliseconds) and Jackson (7.99 milliseconds). This advantage was maintained and increased proportionally with growing file complexities. Memory utilization during deserialization (Figure 5) also demonstrated KotlinX Serialization's efficiency, using significantly less memory on average (24.26 megabytes) compared to Gson (25.52 megabytes) and Jackson (54.78 megabytes). These findings reinforce KotlinX Serialization's optimal memory efficiency, making it particularly suited for large-scale server environments dealing with extensive JSON data.

Regarding CPU consumption during deserialization (depicted on Figure 6), KotlinX Serialization again demonstrated superior efficiency. It maintained the lowest CPU usage at an average of 7.44%, slightly outperforming Gson (7.69%) and Jackson (7.84%). These results underscore KotlinX Serialization's balanced approach to resource utilization, delivering consistent performance advantages across all evaluated metrics.

IV.1.4 Statistical Analysis

Wherever To substantiate the empirical findings, statistical analyses were rigorously conducted. Initial normality testing using the Kolmogorov-Smirnov method indicated that the data sets did not follow a normal distribution. Consequently, non-parametric statistical tests were employed to ensure valid and accurate comparative analyses. The Kruskal-Wallis test, a robust non-parametric method, identified statistically significant differences among the three libraries across all performance metrics and JSON processing tasks. All calculated p-values were significantly below the predetermined alpha level of 0.05, thus confirming meaningful performance differences among Jackson, Gson, and KotlinX Serialization. Post-hoc testing using Dunn's test provided detailed pairwise comparisons, clearly delineating significant performance differences between each library pair. These analyses consistently demonstrated that KotlinX Serialization significantly outperformed Jackson and Gson in execution time and CPU efficiency across parsing and deserialization tasks, while Jackson was superior in memory efficiency during parsing tasks.

IV.1.5 Summary Of Findings

The Overall, the comprehensive evaluation of JSON processing tasks—parsing, serialization, and deserialization—revealed clear and statistically significant performance differences among Jackson, Gson, and KotlinX Serialization. KotlinX Serialization consistently demonstrated superior execution speed and CPU efficiency, particularly pronounced during parsing and deserialization tasks, with moderate yet notable advantages during serialization. Jackson showed distinct strengths in memory utilization, particularly during parsing, and competitive execution speeds for smaller datasets during serialization.

Gson generally performed less effectively relative to Jackson and KotlinX Serialization across all tested metrics, highlighting limitations due to its reflection-based approach. These results collectively indicate that KotlinX Serialization's compile-time code generation strategy significantly enhances performance, making it particularly suitable for performance-critical server applications handling large and complex JSON data. Jackson's reflection-based approach, while efficient in certain scenarios such as memory-constrained environments, exhibits scalability limitations under increasingly complex workloads.

Gson's overall performance disadvantages suggest it may be less suitable for scenarios requiring high computational efficiency or large-scale JSON data handling. In conclusion, this study provides clear, statistically validated empirical evidence highlighting the performance characteristics and trade-offs associated with three prominent JSON libraries in Kotlin server environments. These findings equip developers and architects with actionable insights, enabling informed library selection aligned with specific application requirements and performance objectives.

IV.2 DISCUSSION

IV.2.1 Interpretation Of Findings

The primary objective of this study was to systematically evaluate the comparative performance of three prominent JSON libraries—Jackson, Gson, and KotlinX Serialization—in server-side Kotlin environments. The analysis of experimental data consistently indicated significant performance differences among the tested libraries, each exhibiting distinctive advantages and limitations. KotlinX Serialization emerged prominently as the most efficient library overall, particularly excelling in execution time and CPU consumption during parsing and deserialization tasks. Jackson demonstrated notable strengths in memory utilization and serialization speed for smaller data sets. Conversely, Gson consistently underperformed relative to the other libraries in most performance metrics evaluated.

IV.2.2 Impact of Computational Strategies on Performance

The significant performance differences observed among these libraries can primarily be attributed to their underlying computational strategies. KotlinX Serialization's superior performance, especially regarding execution time and CPU efficiency, can be directly associated with its use of compile-time code generation through compiler plugins. This approach minimizes runtime overhead by avoiding reflection mechanisms commonly employed by Jackson and Gson. The elimination of runtime reflection substantially enhances computational efficiency and reduces resource demands, translating into consistently faster and more efficient data processing. Jackson and Gson rely heavily on runtime reflection, dynamically inspecting and processing data structures at runtime.

While reflection provides flexibility, enabling easy integration across various programming scenarios, it inherently introduces computational overhead and increased resource consumption. This reflection overhead is particularly pronounced in Gson, which consistently showed the lowest performance across execution time, memory usage, and CPU consumption metrics. Jackson, although similarly reliant on reflection, demonstrated superior memory efficiency and competitive execution speeds, particularly in serialization tasks involving smaller datasets. Jackson's performance optimization efforts, such as the use of specialized modules for Kotlin support, likely contribute to its improved performance relative to Gson [23].

IV.2.3 Memory Utilization Analysis

Memory usage results revealed another critical performance dimension. Jackson demonstrated superior memory efficiency during parsing tasks, consistently utilizing significantly less memory compared to Gson and KotlinX Serialization. This efficiency advantage suggests Jackson's reflection-based implementation, augmented by Kotlin-specific optimization modules, effectively manages memory resources during intermediate data processing.

Conversely, Gson's higher memory consumption is likely due to less optimized reflection handling, resulting in increased memory overhead during runtime. In serialization and deserialization tasks, KotlinX Serialization achieved substantial memory efficiency, significantly outperforming both Gson and Jackson. This enhanced efficiency likely results from KotlinX Serialization's compile-time code generation strategy, which streamlines object transformation processes and minimizes runtime memory allocations. This inherent advantage is particularly beneficial in resource-constrained or high-throughput server environments where memory optimization significantly influences overall performance and scalability [12].

IV.2.4 Cpu Consumption and Computational Load

As Analysis of CPU consumption further highlighted KotlinX Serialization's balanced and efficient approach. Consistently recording the lowest CPU usage across all tasks, KotlinX Serialization demonstrated effective computational optimization. The observed lower CPU usage is directly attributable to its compile-time generated serialization logic, reducing runtime complexity and computational demands. In contrast, Jackson and Gson's reflection-based methods inherently increase runtime processing complexity and CPU load, reflecting in consistently higher CPU consumption figures. However, the differences in CPU utilization between Jackson and Gson remained marginal, suggesting similar computational load characteristics inherent to reflection-based libraries.

IV.2.5 Practical Implications for Kotlin-Based Server Applications

From a practical perspective, the findings of this study offer actionable insights into library selection for server-side Kotlin applications. KotlinX Serialization's consistent superiority across multiple performance metrics—execution speed, CPU efficiency, and memory utilization in critical scenarios—makes it the preferred choice for high-performance applications, particularly those requiring rapid, large-scale JSON data processing. Its compile-time code generation strategy aligns closely with Kotlin's native features, ensuring optimal compatibility and performance.

Jackson remains an attractive alternative, particularly in scenarios prioritizing memory efficiency or involving moderate-sized datasets. Its reflection-based approach, supplemented by Kotlin-specific modules, provides sufficient performance for many server-side applications, balancing flexibility and resource usage. In contrast, Gson's lower overall performance metrics suggest limited suitability for demanding or resource-intensive server applications, where efficiency and computational load significantly impact application responsiveness and scalability.

IV.2.6 Study Limitations and Recommendations for Future Research

Footnotes Despite its comprehensive methodological approach, this study has certain limitations warranting consideration in interpreting results and guiding future research. Firstly, the scope was deliberately restricted to evaluating three widely adopted JSON libraries within Kotlin server environments. Future research could expand this scope to include additional libraries or different programming environments to offer broader comparative insights.

Secondly, the experiments were conducted under controlled laboratory conditions using specific hardware and software configurations. While this approach ensures methodological rigor and reproducibility, variations in hardware or deployment environments could influence performance outcomes. Therefore, future studies should consider conducting similar evaluations in diverse real-world deployment contexts to validate and extend these findings.

Additionally, the focus on JSON data with structured nesting and incremental complexity provides relevant insights; however, incorporating other data formats or processing scenarios could further enrich understanding of library performance characteristics. Further research could explore alternative serialization formats or real-world data processing tasks typical of industry-specific applications, thereby broadening the practical relevance and applicability of performance evaluations.

V. CONCLUSIONS

In conclusion, this discussion systematically examined the underlying factors influencing performance differences among Jackson, Gson, and KotlinX Serialization libraries, providing comprehensive interpretations of empirical findings. KotlinX Serialization's compile-time strategy consistently emerged as the most effective approach, delivering superior execution speed, memory efficiency, and CPU optimization suitable for high-performance Kotlin server applications. Jackson's memory efficiency and competitive performance in certain contexts provide valuable alternative options, while Gson's general underperformance highlights limitations inherent to reflection-based processing strategies. These insights enable informed decision-making among developers and system architects, ensuring optimal library selection aligned with specific server-side application requirements.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Achmad Arwan, Julia Nur Fajrina, Eriq Muhammad Adam Jonemaro.
Methodology: Achmad Arwan, Julia Nur Fajrina, Eriq Muhammad Adam Jonemaro.
Investigation: Achmad Arwan, Julia Nur Fajrina, Eriq Muhammad Adam Jonemaro.
Discussion of results: Achmad Arwan, Julia Nur Fajrina, Eriq Muhammad Adam Jonemaro.
Writing – Original Draft: Achmad Arwan, Julia Nur Fajrina, Eriq Muhammad Adam Jonemaro.
Writing – Review and Editing: Achmad Arwan, Julia Nur Fajrina, Eriq Muhammad Adam Jonemaro..
Resources: Achmad Arwan, Julia Nur Fajrina, Eriq Muhammad Adam Jonemaro.
Supervision: Achmad Arwan, Julia Nur Fajrina, Eriq Muhammad Adam Jonemaro.
Approval of the final text: Achmad Arwan, Julia Nur Fajrina, Eriq Muhammad Adam Jonemaro.

VII. REFERENCES

[1] A. Grebenkina, “Ten Years of Kotlin!,” 2021.

[2] JetBrains, “Kotlin - The State of Developer Ecosystem in 2023 Infographic,” 2025.

[3] JetBrains, “Kotlin Serialization Guide,” 2024.

[4] IETF, “RFC 7159 - The JavaScript Object Notation (JSON) Data Interchange Format,” 2014.

[5] H. K. Dhalla, “A Performance Analysis of Native JSON Parsers in Java, Python, MS.NET Core, JavaScript, and PHP,” 16th International Conference on Network and Service Management, CNSM 2020, 2nd International Workshop on Analytics for Service and Application Management, AnServApp 2020 and 1st International Workshop on the Future Evolution of Internet Protocols, IPFutu, 2020, doi: 10.23919/CNSM50824.2020.9269101.

[6] Q. M. o‘g‘li Tohirov, D. A. Ayupova, and R. I. Zakirova, “Exploring Client-Side and Server-Side Architectures in Web Development: A Comprehensive Analysis,” 2024, doi: https://doi.org/10.5281/zenodo.10729101.

[7] M. Zoltu, T. Saloranta, and J. Minard, “Cache reflection results. · Issue #44,” 2016.

[8] J. Friesen, “Parsing and Creating JSON Objects with Gson,” in Java XML and JSON: Document Processing for Java SE, Apress, 2019, pp. 243–298. doi: https://doi.org/10.1007/978-1-4842-4330-5_9.

[9] JetBrains, “Kotlin multiplatform / multi-format serialization,” 2025.

[10] K. Maeda, “Performance evaluation of object serialization libraries in XML, JSON and binary formats,” 2012 2nd International Conference on Digital Information and Communication Technology and its Applications, DICTAP 2012, pp. 177–182, 2012, doi: 10.1109/DICTAP.2012.6215346.

[11] J. Vanura and P. Kriz, Perfomance evaluation of Java, JavaScript and PHP serialization libraries for XML, JSON and binary formats, vol. 10969 LNCS. Springer International Publishing, 2018. doi: 10.1007/978-3-319-94376-3_11.

[12] S. Aigner, “kotlinx.serialization 1.2 Released: High-Speed JSON Handling, Value Class Support, Overhauled Docs, and more,” 2021.

[13] JetBrains, “The State of Developer Ecosystem in 2023,” 2023.

[14] JetBrains, “Kotlin for Server Side,” 2024. [Online]. Available: https://kotlinlang.org/docs/server-overview.html

[15] Kotlin Documentation, “Serialization | Kotlin Documentation,” 2025. [Online]. Available: https://kotlinlang.org/docs/serialization.html

[16] Confluent, “What is Data Serialization? [Beginner’s Guide],” 2024. [Online]. Available: https://www.confluent.io/learn/data-serialization/

[17] B. Kitchenham, S. L. Pfleeger, and D. C. Hoaglin, “Statistical Significance Testing—A Panacea for Software Technology Experiments?,” Journal of Systems and Software, vol. 73, no. 3, pp. 485–501, 2002, doi: 10.1016/j.jss.2004.03.001.

[18] J. Gerrans and R. S. Sherratt, “Comparing XML and JSON Characteristics as Formats for Data Serialization Within Ultralow Power Embedded Systems,” IEEE Embed Syst Lett, vol. 16, no. 4, pp. 489–492, Dec. 2024, doi: 10.1109/LES.2024.3450576.

[19] H. K. Dhalla, “A Performance Analysis of Native JSON Parsers in Java, Python, MS.NET Core, JavaScript, and PHP,” in 2020 16th International Conference on Network and Service Management (CNSM), IEEE, 2020, pp. 1–5. doi: 10.23919/CNSM50824.2020.9269049.

[20] D. P. Proos and N. Carlsson, “Performance Comparison of Messaging Protocols and Serialization Formats for Digital Twins in IoV,” in 2020 IFIP Networking Conference (Networking), 2020, pp. 10–18.

[21] JetBrains, “Get started with Kotlin,” 2025.

[22] K. M. Ramachandran and C. P. Tsokos, “Chapter 12 - Nonparametric Tests,” in Mathematical Statistics with Applications in R (Second Edition), Academic Press, 2015, pp. 589–637. doi: https://doi.org/10.1016/B978-0-12-417113-8.00012-6.

[23] FasterXML, “Jackson Module Kotlin,” 2025.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 728-737. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3048

RESEARCH ARTICLE OPEN ACCESS

TOWARD ETHICAL AND INTELLIGENT SENTIMENT ANALYSIS IN TOURISM: FUTURE DIRECTIONS FOR THE NEXT TWO DECADES

Hardian Oktavianto[*1], Ika Safitri Windiarti[2] and Bagus Setya Rintyarna[3]

[1, 3]Universitas Muhammadiyah Jember
[2]Universiti Muhammadiyah Malaysia

[1]https://orcid.org/0000-0002-2446-7401, [2]https://orcid.org/0000-0002-9603-3397, [3]https://orcid.org/0000-0002-3275-0995,

Email: *hardian@unmuhjember.ac.id, ika.windiarti@umam.edu.my, bagus.setya@unmuhjember.ac.id

ARTICLE INFO

Article History
Received: December 4, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Sentiment analysis,
Tourism analystics,
Ethical governance,
Cultural intelligence.

ABSTRACT

Tourism in the digital era is increasingly influenced by user-generated content, where sentiment analysis plays a critical role in understanding traveler perceptions and guiding destination management. This study reviews advancements in sentiment analysis from 2015 to 2024 and forecasts its evolution through 2045, emphasizing the integration of Artificial Intelligence, Natural Language Processing, and ethical-cultural awareness. Employing a narrative literature review, the research synthesizes over 120 academic works to identify technological, ethical, and linguistic challenges shaping the field. The paper proposes the Human-AI Integrative Sentiment Framework (HAISF), a holistic model uniting five dimensions: data diversity, computational intelligence, ethical governance, cultural sensitivity, and sustainable application. This framework supports creating sentiment analysis systems that are easy to understand, protect user privacy, and respect cultural differences, ensuring that the use of artificial intelligence stays consistent with human values. The study concludes that future sentiment analysis in tourism will evolve toward an empathic, multimodal, and ethically grounded paradigm, enabling predictive and socially responsible decision-making across global tourism ecosystems.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Tourism is increasingly shaped by digital communication, where travelers continuously express opinions through social media, online reviews, and other user-generated content (UGC) [1]. Understanding these sentiments has become vital for destinations and tourism businesses seeking to improve customer experiences and manage brand reputation effectively [2]. Sentiment analysis, an interdisciplinary approach combining computational linguistics, machine learning, and data analytics, has thus emerged as an indispensable tool for extracting meaningful insights from large volumes of textual and visual data [3]. Early research in tourism sentiment analysis primarily focused on polarity detection (positive, neutral, negative) using lexicon-based and classical machine learning models such as Support Vector Machines and Naïve Bayes [4].

Recent advances in deep learning and transformer architectures, including BERT and RoBERTa, have markedly enhanced the precision of sentiment classification by capturing semantic and contextual nuances [5]. Moreover, multimodal sentiment analysis that integrates text, image, and audio inputs offers a more holistic understanding of tourist experiences [6]. Despite these advances, several challenges persist. Existing models often fail to account for linguistic and cultural diversity, leading to biased interpretations of sentiment [7]. Ethical and privacy concerns also arise as AI systems increasingly rely on personal data. Furthermore, current research remains fragmented, lacking a unified framework that connects technological progress with ethical and socio-cultural considerations in tourism. To address these gaps, this paper aims to forecast the future of sentiment analysis in tourism over the next 20 years, focusing on three objectives:

1. To review and synthesize recent developments in sentiment analysis relevant to tourism;

2. To identify emerging trends and challenges that will shape the field through 2045; and
3. To propose a conceptual framework integrating technology, ethics, and inclusivity for future tourism analytics.

The remainder of this paper is structured as follows: Section 2 describes the research approach; Section 3 reviews the evolution and current state of sentiment analysis in tourism; Section 4 outlines emerging trends and projections; Section 5 discusses ethical and linguistic challenges; Section 6 presents a conceptual framework for future research; and Section 7 concludes with implications for academia and industry.]

II. METHODOLOGY

This study adopts a narrative literature review approach to synthesize academic and applied research on sentiment analysis in tourism from 2015 to 2024. Unlike systematic reviews that focus on quantitative inclusion criteria, a narrative review allows for critical interpretation and theoretical integration across multidisciplinary studies [8]. This approach is particularly appropriate given the rapid evolution of Artificial Intelligence (AI), Natural Language Processing (NLP), and data analytics in tourism research. The literature was retrieved from major academic databases, Scopus, Web of Science, IEEE Xplore, ScienceDirect, and SpringerLink, using a combination of keywords: “sentiment analysis,” “tourism analytics,” “machine learning,” “deep learning,” “natural language processing,” and “AI ethics”. Peer-reviewed journal articles, conference papers, and review studies were included, while opinion pieces, editorials, and non-English publications were excluded. In total, over 120 recent publications were reviewed, with a focus on methodological innovation, application context, and theoretical contribution. The analysis process followed three stages:

1. Conceptual mapping: Identification of key research themes and technological milestones in sentiment analysis within tourism.
2. Thematic synthesis: Grouping studies into categories, methodological development, current applications, ethical/cultural challenges, and future trends.
3. Gap identification and forecasting: Extraction of research gaps and projection of future directions for the next two decades based on technological trajectories and policy reports.

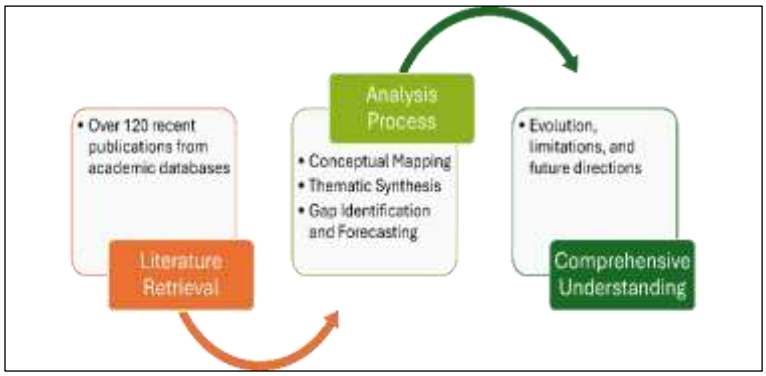


Figure 1: Method.
Source: Authors, (2026).

This process ensured a comprehensive understanding of how sentiment analysis has evolved, the limitations of existing research, and the directions for its future integration into the tourism domain.

III. EVOLUTION AND CURRENT STATE OF SENTIMENT ANALYSIS IN TOURISM

III.1 EARLY DEVELOPMENT AND TECHNOLOGICAL MILESTONES

Sentiment analysis has evolved from simple text classification into a sophisticated, multi-layered analytical process integrating AI and deep learning. Early applications in tourism primarily relied on lexicon-based and machine learning approaches to determine polarity (positive, negative, neutral) in user reviews [4]. Traditional algorithms such as Naïve Bayes [9], Support Vector Machines (SVM), and Decision Trees demonstrated moderate success but struggled with contextual nuances like sarcasm, idioms, and cultural expressions [10]. From 2015 to 2018, Aspect-Based Sentiment Analysis (ABSA) gained attention for its ability to link sentiments to specific service components, such as hotel cleanliness, transportation, or food quality, providing richer managerial insights [11]. The introduction of deep learning architectures, notably Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), marked a methodological leap, allowing models to capture sequential and semantic dependencies within text data [5].

III.2 THE DEEP LEARNING AND TRANSFORMER REVOLUTION

The period after 2019 witnessed a paradigm shift with the introduction of transformer-based language models, including BERT, RoBERTa, and GPT architectures. These models significantly improved accuracy by capturing contextual meaning across entire sentences and multilingual corpora [12]. Such advances enabled researchers to analyze vast and heterogeneous datasets, from TripAdvisor and Yelp reviews to multimodal inputs like photos and videos. Recent studies in tourism have leveraged multimodal sentiment analysis (MSA) to combine textual, visual, and acoustic data, leading to a more comprehensive understanding of traveler emotions [6]. For example, photo-based sentiment analysis in dark tourism sites has revealed that negative emotions often dominate affective responses, offering insights into destination image management [13]. Similarly, transformer-enhanced systems like ChatGPT-based models have demonstrated superior interpretability and accuracy compared with conventional machine learning frameworks [14].

III.3 CURRENT APPLICATIONS IN THE TOURISM INDUSTRY

Sentiment analysis has become a cornerstone of destination management, service quality assessment, and customer relationship management (CRM) in tourism. Organizations increasingly rely on real-time sentiment tracking to evaluate traveler satisfaction, forecast market trends, and tailor marketing strategies [4], [15]. In hospitality, for instance, automated analysis of guest reviews informs operational improvements and service personalization. Airlines have applied sentiment analysis to monitor customer feedback on social media, achieving up to 90% accuracy in classifying sentiment using hybrid BERT–fastText models [16].
Furthermore, Big Data and tourism intelligence systems now incorporate AI-driven sentiment analytics to identify emerging travel patterns, predict seasonal preferences, and manage brand reputation. Platforms such as ReUS [17] and BERTopic-based frameworks [18] demonstrate how unsupervised and topic-modeling methods enhance interpretability for decision-makers. These developments collectively indicate that sentiment analysis is not merely a monitoring tool but a strategic asset driving innovation, competitiveness, and sustainability in the tourism sector.

Table 1: Evolution And Current State Of Sentiment Analysis In Tourism.

Period / Phase	Technological Milestones	Key Methods & Models	Application Contexts in Tourism	Limitations / Challenges	Representative Outcomes or Contributions
Early Development (Pre-2015)	Emergence of sentiment analysis in tourism using basic text classification	Lexicon-based methods; Machine Learning (Naïve Bayes, SVM, Decision Trees)	Analysis of online reviews (TripAdvisor, Booking.com) to determine polarity (positive, neutral, negative)	Inability to handle sarcasm, idioms, and cultural nuances	Provided initial automation of opinion mining and polarity detection in tourism reviews
Aspect-Based Expansion (2015–2018)	Integration of Aspect-Based Sentiment Analysis (ABSA) for fine-grained insights	ABSA frameworks; feature engineering techniques	Evaluation of specific service attributes (e.g., hotel cleanliness, transport, food quality)	Limited scalability; dependence on manual feature extraction	Enabled tourism managers to link sentiments to concrete service dimensions
Deep Learning Era (2018–2019)	Adoption of neural network models for contextual sentiment understanding	CNNs, RNNs, LSTM-based models	Automated analysis of complex textual data from online travel platforms	High computational cost; limited interpretability	Enhanced semantic and sequential text understanding, improving sentiment classification accuracy
Transformer Revolution (2019–2022)	Introduction of transformer-based NLP architectures	BERT, RoBERTa, GPT models	Multilingual and context-aware sentiment analysis; large-scale tourism datasets (TripAdvisor, Yelp)	Requires large labeled data; risk of bias in pretrained models	Achieved state-of-the-art accuracy; improved cross-lingual and contextual sentiment interpretation
Multimodal & Hybrid Systems (2022–2024)	Expansion toward multimodal and hybrid AI-driven sentiment systems	MSA (text + image + audio); hybrid BERT–fastText; topic modeling (BERTopic)	Analysis of traveler emotions via text, photos, and videos; CRM and destination image management	Data integration complexity; privacy and ethical considerations	Enabled real-time tourism intelligence, trend prediction, and customer experience optimization

Source: Authors, (2026).

The evolution of sentiment analysis in tourism has progressed from rule-based polarity detection to sophisticated AI-driven multimodal frameworks. These developments have transformed sentiment analysis from a descriptive tool into a strategic asset for data-driven decision-making, competitiveness, and sustainable tourism management.

IV. EMERGING TRENDS AND FUTURE DIRECTIONS (2025–2045)

The next two decades are expected to mark a transformative era in sentiment analysis, driven by accelerated developments in artificial intelligence, machine learning, and natural language processing [19]. These advancements will not only enhance the precision and interpretability of sentiment classification but will also expand its applicability across the tourism value chain, from destination management and service personalization to predictive marketing and sustainability analytics. The following subsections outline key technological, methodological, and interdisciplinary trends anticipated to define the field through 2045.

IV.1 THE RISE OF LARGE LANGUAGE MODELS AND GENERATIVE AI

The proliferation of Large Language Models (LLMs), such as GPT, LLaMA, and Claude, has already redefined text-based analytics, enabling systems to understand and generate human-like responses with remarkable contextual accuracy. Over the next two decades, sentiment analysis will likely integrate domain-adapted LLMs trained specifically on tourism-related data (e.g., hotel reviews, travel blogs, and destination narratives). These models will outperform general-purpose systems by capturing tourism-specific expressions, idioms, and cultural nuances.

Emerging frameworks will combine neurosymbolic AI, which integrates neural networks with symbolic reasoning [20], to improve explainability and ethical accountability in sentiment decisions. Moreover, fine-tuning techniques and transfer learning will enable models to adapt to new linguistic and cultural contexts with minimal retraining effort, facilitating cross-regional tourism analytics. By 2045, generative AI systems will not only analyze but also simulate tourist sentiments, predicting how policy changes, marketing campaigns, or sustainability initiatives might shape public perceptions before they occur.

IV.2 MULTIMODAL AND REAL-TIME SENTIMENT ANALYSIS

Future sentiment analysis will increasingly move beyond text, leveraging multimodal learning to integrate data from images, videos, voice recordings, and physiological sensors. Advances in multimodal architectures, such as Text-Centric Sharing-Private (TCSP) and Modality Translation-based Multimodal Sentiment Analysis (MTMSA) models, have demonstrated the ability to capture complex emotional cues even when certain modalities are missing [21], [22]. In the context of tourism, multimodal analysis could synthesize visual impressions (e.g., Instagram photos), acoustic tones (e.g., traveler vlogs), and textual reviews to construct a 360-degree emotional map of tourist experiences. Combined with Internet of Things (IoT) sensors embedded in smart destinations, such systems could enable real-time mood analytics, offering immediate feedback to tourism operators about visitor satisfaction, crowd sentiment, and service performance. By 2040, real-time sentiment dashboards powered by edge computing and 6G connectivity could allow destination managers to monitor emotional dynamics across attractions, ensuring responsive service design and crisis management.

IV.3 CROSS-LINGUAL AND CULTURALLY ADAPTIVE SYSTEMS

Language and cultural diversity remain among the greatest challenges in global tourism sentiment analysis. The coming years will see widespread deployment of Cross-Lingual Sentiment Analysis (CLSA) models, which use machine translation, multilingual embeddings, and zero-shot transfer learning to interpret sentiments across languages with limited labeled data [23]. Research has shown that fine-tuned models trained on high-resource languages like English can achieve up to 96% accuracy when adapted to low-resource languages such as Arabic or Tamil [24], [25]. Beyond linguistic translation, future sentiment systems must become culturally adaptive, incorporating local emotional semantics and indigenous knowledge systems. This shift requires interdisciplinary collaboration among linguists, anthropologists, and computer scientists to ensure that AI models reflect diverse worldviews rather than reproducing Western-centric emotion taxonomies. Such inclusivity will be essential for creating equitable, context-aware sentiment analysis tools that resonate with multicultural audiences.

IV.4 INTEGRATION WITH PREDICTIVE ANALYTICS AND DECISION INTELLIGENCE

Between 2030 and 2045, sentiment analysis will evolve from descriptive evaluation toward predictive and prescriptive intelligence. By combining emotional data with behavioral, transactional, and geospatial information, AI systems will forecast tourism demand patterns, destination reputation risks, and visitor satisfaction trends. Techniques such as probabilistic linguistic term sets, multi-granularity decision systems, and emotion variance analysis (EVA) are expected to enhance the interpretive granularity of these predictions [26], [27]. Integrated with Big Data analytics and decision-support systems, these models will enable policymakers and business leaders to conduct scenario planning, for example, anticipating the emotional impacts of climate change, political unrest, or pandemics on tourism sentiment.

IV.5 TOWARD ETHICAL, TRANSPARENT, AND SUSTAINABLE AI

As AI becomes more pervasive, ethical governance will be central to its sustainable deployment. By 2035, sentiment analysis systems are expected to adopt explainable AI (XAI) frameworks that make decision logic transparent to both researchers and end-users. Parallelly, privacy-preserving machine learning techniques, such as federated learning and differential privacy, will become standard to ensure data protection without compromising analytic performance [28]. Regulatory harmonization across jurisdictions, building on frameworks such as the EU General Data Protection Regulation (GDPR) and emerging AI governance acts, will dictate responsible data collection, model training, and sentiment interpretation practices. These ethical innovations will not only mitigate bias and misuse but also strengthen public trust in AI-driven tourism analytics.

IV.6 INTERDISCIPLINARY CONVERGENCE AND THE HUMAN-AI COLLABORATION PARADIGM

The future of sentiment analysis lies in interdisciplinary convergence, where advances in computer science intersect with behavioral psychology, linguistics, and tourism studies [29]. Human-AI collaboration will redefine how emotional intelligence is operationalized in digital tourism environments. AI agents will increasingly act as empathic interfaces, capable of detecting and responding to human emotions in customer service, destination management, and online communication. This will create new roles for human experts, such as AI ethics auditors, data curators, and emotion designers, who oversee fairness, transparency, and inclusivity in AI systems. By integrating human empathy with machine precision, the tourism sector can foster more personalized, ethical, and emotionally resonant experiences.

IV.7 SUMMARY OF FUTURE DIRECTIONS

The emerging trends in sentiment analysis for tourism highlight a shift toward more intelligent, inclusive, and human-centered systems. Large Language Models (LLMs), such as domain-specific GPT variants, enable real-time interpretation of tourist feedback and predictive insight generation. Multimodal analytics, combining text, image, and voice data, provides a holistic view of travelers' emotions and destination experiences. Cross-lingual adaptation through multilingual transfer learning promotes inclusivity and expands global market intelligence.

Predictive and prescriptive AI systems leverage Big Data to forecast emotions and support proactive decision-making, including crisis management. Meanwhile, the rise of ethical and explainable AI ensures transparency, privacy, and compliance in analytical ecosystems. Finally, human–AI collaboration merges automation with empathy, fostering more personalized engagement and enhancing overall visitor satisfaction in the tourism sector.

Table 2: Summary of Emerging Trends and Future Directions.

Trend	Description	Tourism Implication
Large Language Models (LLMs)	Domain-specific, context-aware models (e.g., GPT variants)	Real-time analysis of tourist reviews and predictive insights
Multimodal Analytics	Integration of text, image, and voice sentiment	Comprehensive emotional mapping of destinations
Cross-Lingual Adaptation	Multilingual transfer learning and inclusivity	Broader market intelligence and global reach
Predictive & Prescriptive Systems	Emotion forecasting via AI and Big Data	Proactive decision-making and crisis anticipation
Ethical & Explainable AI	Transparent, privacy-preserving models	Trustworthy and compliant analytics ecosystems
Human–AI Collaboration	Integration of empathy with automation	Enhanced visitor engagement and satisfaction

Source: Authors, (2026).

V. ETHICAL, CULTURAL, AND LINGUISTIC CHALLENGES

As sentiment analysis becomes increasingly integrated into tourism management and policymaking, ethical, cultural, and linguistic considerations will play a defining role in shaping its credibility and societal impact. While advanced algorithms such as deep learning and Large Language Models have greatly improved accuracy and scalability, they also introduce challenges related to data bias, privacy, and cross-cultural misrepresentation. Addressing these concerns is essential to ensure that sentiment analysis remains transparent, inclusive, and responsible in its application within the tourism sector.

V.1 ALGORITHMIC BIAS AND CULTURAL REPRESENTATION

A major ethical issue in sentiment analysis arises from algorithmic bias, which occurs when AI models inherit distortions embedded in their training data. Most sentiment analysis systems are trained predominantly on English-language and Western datasets, resulting in models that misinterpret expressions, emotions, or sarcasm from other cultural contexts [30]. This bias can distort perceptions of destinations or communities, particularly in multicultural and indigenous tourism settings, where local emotional expressions may diverge significantly from standardized sentiment lexicons. To mitigate such risks, future research must emphasize culturally sensitive data curation and diversity-aware training frameworks. Approaches grounded in feminist epistemology and critical race theory [31] provide valuable foundations for rethinking how cultural perspectives are represented in AI systems. Collaborative initiatives that involve local experts and community stakeholders can help develop contextual sentiment lexicons and emotion ontologies that accurately capture non-Western affective nuances.

V.2 LINGUISTIC DIVERSITY AND LOW-RESOURCE LANGUAGE CHALLENGES

Tourism operates across linguistic boundaries, yet most sentiment models remain heavily optimized for high-resource languages such as English, Mandarin, or Spanish. Low-resource languages—including those spoken in Southeast Asia, Africa, and the Pacific Islands—suffer from limited annotated data and linguistic resources, reducing the inclusivity of global sentiment analysis [23]. Recent progress in Cross-Lingual Sentiment Analysis (CLSA) and transfer learning has shown promise, enabling models trained on one language to perform reasonably well on others using multilingual embeddings or machine translation [25].

However, such systems still risk losing cultural and emotional nuances during translation, particularly when dealing with idiomatic or context-specific expressions. To address this limitation, the development of localized NLP models that incorporate region-specific datasets and hybrid lexicons is crucial. The creation of multimodal datasets, combining visual, auditory, and textual cues, can further compensate for linguistic gaps by leveraging non-verbal indicators of emotion. Ultimately, linguistic inclusivity is not merely a technical necessity but an ethical imperative, ensuring that all voices in global tourism are accurately represented.

V.3 DATA PRIVACY, CONSENT, AND TRANSPARENCY

The collection and analysis of personal data, such as online reviews, social media posts, and biometric inputs, raise serious data privacy and consent issues. Many AI systems used for sentiment analysis process user-generated content without explicit permission, challenging international privacy frameworks such as the General Data Protection Regulation (GDPR) and emerging AI governance acts [28]. Beyond legal compliance, there are ethical concerns about how this data is used, shared, and potentially monetized. The use of commercial AI platforms for qualitative analysis introduces further complications, as proprietary algorithms may lack transparency and accountability [32].

Moreover, anonymization is often insufficient when datasets include geotagged or image-based tourism content that can re-identify individuals. Future sentiment analysis frameworks must therefore adopt privacy-preserving machine learning methods such as federated learning, which allows models to learn from distributed data without centralizing it, and differential privacy, which adds statistical noise to protect sensitive information [33]. Ethical data governance must also include clear consent protocols, traceability of data sources, and public transparency reports to build user trust and compliance with global regulatory standards.

V.4 ETHICAL ACCOUNTABILITY AND EXPLAINABLE AI

As AI-driven analytics increasingly influence tourism decision-making, ensuring ethical accountability and explainability becomes paramount. Current deep learning models are often considered "black boxes," making it difficult to interpret how sentiment scores are derived [34]. This lack of transparency undermines stakeholder confidence, especially when sentiment insights inform marketing campaigns, policy design, or destination reputation management. The adoption of Explainable AI (XAI) frameworks offers a promising pathway toward greater interpretability. XAI can make model behavior more understandable by highlighting which linguistic or visual features contributed most to a sentiment classification. Such transparency is critical in tourism contexts where decisions can directly affect public perception, economic outcomes, and community well-being. Ethical auditing of AI systems, through interdisciplinary teams of data scientists, ethicists, and tourism experts, should become standard practice by the 2030s to ensure compliance with both technical accuracy and moral responsibility.

V.5 THE HUMAN DIMENSION: EMOTIONAL LABOR AND AI CO-CREATION

An often-overlooked aspect of sentiment analysis is the human labor involved in data annotation, validation, and interpretation. The reliance on poorly compensated annotators for labeling emotional content raises concerns about digital labor ethics, particularly in developing economies. Furthermore, excessive automation risks eroding the human empathy required in interpreting affective data, especially in service-oriented industries like tourism. Looking ahead, the focus should shift from replacing human judgment to co-creating human–AI partnerships that enhance ethical and cultural understanding. AI systems designed to assist rather than replace human evaluators can combine computational scale with emotional intelligence, ensuring that technological progress supports social and ethical tourism development.

V.6 SUMMARY OF ETHICAL, CULTURAL, AND LINGUISTIC CHALLENGES

Table 3: Summary Of Challenges in Ethical, Cultural, And Linguistic.

Challenge	Key Issue	Recommended Mitigation Strategy
Algorithmic Bias	Overrepresentation of Western data leading to cultural misinterpretation	Culturally diverse datasets; inclusion of indigenous lexicons
Linguistic Inequality	Limited support for low-resource languages	Cross-lingual transfer learning and multimodal data integration
Privacy and Consent	Unauthorized use of personal data	Federated learning, differential privacy, transparent consent mechanisms
Lack of Explainability	Black-box AI models in tourism analytics	Explainable AI (XAI) and ethical audit frameworks
Human Displacement	Over-automation of emotional interpretation	Human–AI collaboration emphasizing empathy and inclusivity

Source: Authors, (2026).

V.7 TOWARD RESPONSIBLE AND INCLUSIVE AI IN TOURISM

Ethical and linguistic inclusivity is not merely a technical requirement but a cornerstone of sustainable tourism analytics. Future sentiment analysis models must prioritize fairness, transparency, and contextual sensitivity to ensure that AI-driven insights contribute positively to human-centered tourism development. By embedding ethical reflexivity and cross-cultural understanding into their design, these systems can strengthen public trust, support equitable participation, and advance the responsible digital transformation of the global tourism industry.

VI. CONCEPTUAL FRAMEWORK FOR FUTURE RESEARCH

The rapid technological, ethical, and cultural shifts outlined in preceding sections indicate that sentiment analysis in tourism will evolve toward a holistic, human-centered paradigm. To guide this transition, this study proposes a conceptual framework for next-generation sentiment analysis in tourism, referred to as the Human, AI Integrative Sentiment Framework (HAISF). This framework synthesizes technological innovation with ethical governance and cultural inclusivity, offering a roadmap for future research and practical implementation.

VI.1 RATIONALE FOR THE FRAMEWORK

The evolution of sentiment analysis has thus far been driven primarily by technological capability, machine learning algorithms, deep learning architectures, and large-scale data analytics. However, this progress has often outpaced consideration of ethical alignment, interpretability, and cross-cultural representation. The HAISF model addresses this imbalance by positioning technology, ethics, and human–cultural context as interdependent pillars of sustainable AI development for the tourism domain. This integration ensures that sentiment analysis systems not only enhance predictive accuracy but also reflect the emotional diversity and ethical integrity necessary for trustworthy AI in global tourism applications.

VI.2 COMPONENTS OF THE HUMAN–AI INTEGRATIVE SENTIMENT FRAMEWORK (HAISF)

The HAISF consists of five interconnected layers, each representing a critical dimension in future sentiment analysis research and practice.

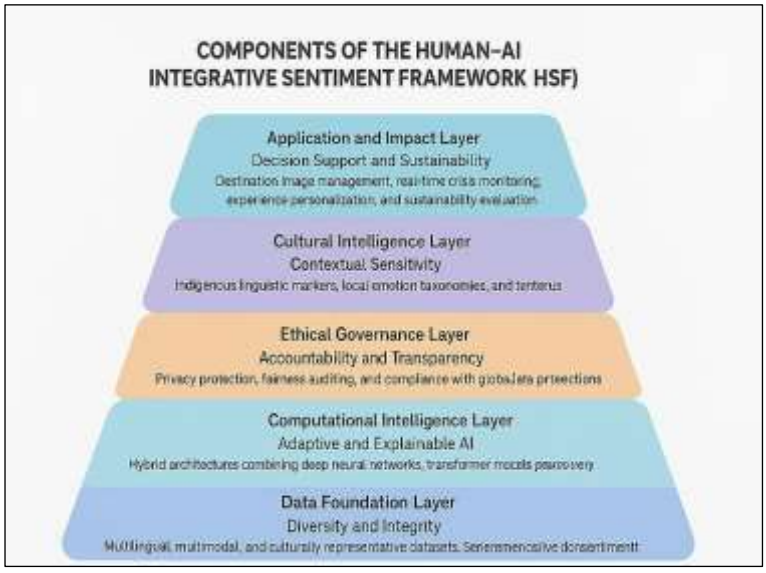


Figure 2: Components of HAISF.
Source: Authors, (2026).

1. Data Foundation Layer: Diversity and Integrity

This layer emphasizes the importance of multilingual, multimodal, and culturally representative datasets. Future models should integrate data from text, images, voice, and sensor inputs while ensuring data authenticity and privacy compliance. Federated data collection methods can preserve user anonymity while supporting continuous learning across decentralized sources.

2. Computational Intelligence Layer: Adaptive and Explainable AI

At the core of the framework lies the computational engine powered by hybrid AI architectures, combining deep neural networks, transformer models (e.g., GPT, BERT, RoBERTa), and neurosymbolic reasoning.

This layer prioritizes explainability and interpretability, enabling stakeholders to understand how sentiment decisions are made. By adopting Explainable AI (XAI) and neuro-symbolic logic, this component ensures transparent, bias-aware sentiment outcomes.

3. Ethical Governance Layer: Accountability and Transparency

This layer operationalizes ethical AI principles by embedding privacy protection, fairness auditing, and accountability mechanisms into sentiment workflows. Ethical governance includes compliance with global data protection standards (e.g., GDPR, ISO/IEC 42001), ethical auditing boards, and open-source algorithmic transparency.

Tourism organizations can employ this layer as a responsibility checkpoint, ensuring data-driven decisions align with social and moral standards.

4. Cultural Intelligence Layer: Contextual Sensitivity

This layer bridges computational and human understanding by embedding local emotion taxonomies, indigenous linguistic markers, and culturally adaptive lexicons into sentiment analysis.

Collaboration between data scientists, linguists, and tourism scholars will allow AI models to interpret affective expressions within their correct socio-cultural frames, reducing misclassification and reinforcing inclusivity in global tourism analytics.

5. Application and Impact Layer: Decision Support and Sustainability

The top layer focuses on the translation of analytical results into strategic decision-making. It includes applications such as destination image management, real-time crisis monitoring, experience personalization, and sustainability evaluation.

By integrating predictive sentiment trends with behavioral and geospatial analytics, this layer empowers policymakers and tourism businesses to act proactively rather than reactively.

VI.3 FUNCTIONAL DYNAMICS OF THE FRAMEWORK

The five layers of HAISF interact dynamically rather than hierarchically. Data flows upward, from raw multimodal input to actionable insights, while ethical and cultural feedback mechanisms flow downward to recalibrate algorithmic behavior. For example, when a system identifies a bias in sentiment interpretation (e.g., misclassification of cultural idioms), the Ethical and Cultural layers can trigger adaptive retraining within the Computational layer, leading to iterative model improvement. This feedback-oriented architecture ensures continuous alignment between technological advancement and social responsibility.

VI.4 RESEARCH DIRECTIONS EMERGING FROM THE FRAMEWORK

The HAISF framework highlights several key research pathways for the next generation of sentiment analysis in tourism. These research directions as shown in table 4, collectively illustrate the interdisciplinary nature of future sentiment analysis, spanning computer science, social ethics, and tourism management.

Table 4: Summary of Research Directions.

Research Area	Proposed Direction	Expected Contribution
Multimodal Sentiment Fusion	Combine text, image, audio, and sensor data into unified affective models	Enhanced accuracy and contextual depth in tourism experience analysis
Cross-Lingual and Cultural Modelling	Develop culturally adaptive emotion lexicons and multilingual AI models	Broader global applicability and inclusivity
Explainable and Trustworthy AI	Create interpretable sentiment systems through XAI and neuro-symbolic methods	Increased stakeholder confidence and ethical compliance
Privacy-Preserving Learning	Implement federated and differential privacy algorithms	Ethical handling of sensitive user-generated data
Predictive Decision Intelligence	Integrate sentiment trends with economic and behavioral forecasting models	Proactive destination management and crisis prediction
Human–AI Collaboration Studies	Explore how humans and AI co-create insights in tourism contexts	Balanced human judgment with machine scalability

Source: Authors, (2026).

VI.5 THEORETICAL AND PRACTICAL IMPLICATIONS

From a theoretical perspective, the HAISF advances current understanding by framing sentiment analysis not only as a computational task but as a socio-technical ecosystem shaped by human values, language, and culture. It encourages tourism scholars to integrate affective computing theory, data ethics, and cross-cultural communication within analytical models.

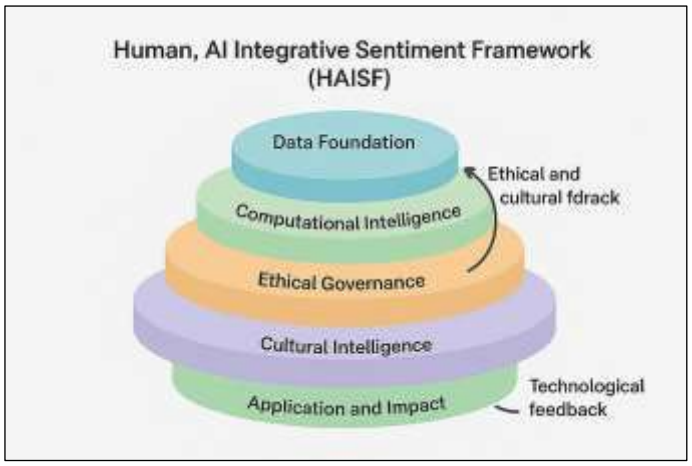


Figure 3: HAISF Framework.
Source: Authors, (2026).

Practically, the framework offers tourism organizations a blueprint for developing AI-driven sentiment systems that are accurate, explainable, and socially responsible. When implemented, HAISF can enhance tourist satisfaction analytics, support evidence-based marketing, and promote sustainable destination management through emotionally intelligent decision-making.

VII. CONCLUSIONS

This study presents the evolution of sentiment analysis in tourism from 2025 to 2045, emphasizing the convergence of Artificial Intelligence, Natural Language Processing, and ethical-cultural awareness. It highlights that sentiment analysis has advanced beyond basic polarity detection to become multimodal and context-sensitive, capable of interpreting emotions through text, image, and voice across diverse linguistic and cultural landscapes. Despite these advancements, ongoing challenges such as algorithmic bias, linguistic inequality, and privacy concerns demand a more balanced and responsible approach. To address these issues, the research introduces the Human-AI Integrative Sentiment Framework (HAISF), a holistic model that integrates technological innovation, ethical governance, and cultural intelligence. The framework reframes sentiment analysis as a socio-technical system, emphasizing adaptive feedback loops between human values and AI-driven processes. This theoretical shift contributes to the development of responsible and inclusive AI in tourism analytics.

Practically, the HAISF provides a roadmap for tourism organizations to develop explainable, privacy-preserving, and ethically grounded sentiment systems. By adopting such systems, stakeholders can gain real-time emotional insights, enhance marketing and service personalization, ensure ethical compliance, and support sustainable destination management. Future research directions include the creation of multimodal and cross-cultural datasets, the advancement of explainable AI, the development of emotion taxonomies sensitive to cultural diversity, and the exploration of privacy-preserving data governance. Additionally, integrating human empathy with AI-driven insights and employing predictive emotional forecasting will further refine tourism intelligence. The future of sentiment analysis in tourism lies in integrating human values, ethical principles, and cultural context with computational intelligence. Through this alignment, sentiment analysis can evolve into a predictive, empathic, and socially responsible decision-support system, driving innovation, trust, and sustainability in the global tourism ecosystem.

VIII. AUTHOR'S CONTRIBUTION

Conceptualization: Hardian Oktavianto, Ika Safitri Windiarti
Methodology: Hardian Oktavianto, Ika Safitri Windiarti, Bagus Setya Rintyarna.
Investigation: Hardian Oktavianto, Ika Safitri Windiarti, Bagus Setya Rintyarna.
Discussion of results: Hardian Oktavianto, Ika Safitri Windiarti, Bagus Setya Rintyarna.
Writing – Original Draft: Hardian Oktavianto.
Writing – Review and Editing: Hardian Oktavianto, Ika Safitri Windiarti, Bagus Setya Rintyarna.
Resources: Hardian Oktavianto, Ika Safitri Windiarti, Bagus Setya Rintyarna.
Supervision: Ika Safitri Windiarti, Bagus Setya Rintyarna.
Approval of the final text: Ika Safitri Windiarti, Bagus Setya Rintyarna.

IX. REFERENCES

[1] M. A. Asyraff, M. H. Hanafiah, N. A. M. Zain, and D. Hariani, "Unboxing the paradox of social media user-generated content (UGC) information qualities and tourist behaviour: moderating effect of perceived travel risk," J. Hosp. Tour. Insights, vol. 7, no. 4, pp. 1809–1830, Jul. 2023, doi: 10.1108/JHTI-02-2023-0072.

[2] K. L. Tan, C. P. Lee, and K. M. Lim, "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research," Appl. Sci., vol. 13, no. 7, 2023, doi: 10.3390/app13074550.

[3] U. Singh, K. Abhishek, and H. K. Azad, "A Survey of Cutting-edge Multimodal Sentiment Analysis," ACM Comput. Surv., vol. 56, no. 9, Apr. 2024, doi: 10.1145/3652149.

[4] A. R. Alaei, S. Becken, and B. Stantic, "Sentiment Analysis in Tourism: Capitalizing on Big Data," J. Travel Res., vol. 58, no. 2, pp. 175–191, 2019, doi: 10.1177/0047287517747753.

[5] T. Abdullah and A. Ahmet, "Deep Learning in Sentiment Analysis: Recent Architectures," ACM Comput. Surv., vol. 55, no. 8, Dec. 2022, doi: 10.1145/3548772.

[6] Q. Lu, X. Sun, Y. Long, Z. Gao, J. Feng, and T. Sun, "Sentiment Analysis: Comprehensive Reviews, Recent Advances, and Open Challenges," IEEE Trans. Neural Networks Learn. Syst., pp. 1–21, 2023, doi: 10.1109/TNNLS.2023.3294810.

[7] J. P. Venugopal, A. A. Subramanian, G. Sundaram, M. Rivera, and P. Wheeler, "A Comprehensive Approach to Bias Mitigation for Sentiment Analysis of Social Media Data," Applied Sciences, vol. 14, no. 23. 2024. doi: 10.3390/app142311471.

[8] S. Hall and E. Leeder, "Narrative reanalysis: A methodological framework for a new brand of reviews," Res. Synth. Methods, vol. 15, no. 6, pp. 1017–1030, Nov. 2024, doi: https://doi.org/10.1002/jrsm.1751.

[9] H. Oktavianto, Q. Ayun, H. A. Al Faruq, L. Handayani, D. Lusiana, and M. F. A. H. T. B. Aj, "The Effect of Laplace Smoothing on Naive Bayes for Sentiment Analysis of the Impact of the AstraZeneca Vaccine," in 2024 Beyond Technology Summit on Informatics International Conference (BTS-I2C), 2024, pp. 625–630. doi: 10.1109/BTS-I2C63534.2024.10941727.

[10] A. Pandey and D. K. Vishwakarma, "Progress, Achievements, and Challenges in Multimodal Sentiment Analysis Using Deep Learning: A Survey," Appl. Soft Comput., 2023, doi: 10.1016/j.asoc.2023.111206.

[11] A. Ameur, S. Hamdi, and S. Ben Yahia, "Sentiment Analysis for Hotel Reviews: A Systematic Literature Review," ACM Comput. Surv., vol. 56, no. 2, Sep. 2023, doi: 10.1145/3605152.

[12] M. S. Islam, M. N. Kabir, N. A. Ghani, K. Z. Zamli, N. S. A. Zulkifli, and M. A. Moni, "'Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach,'" Artif. Intell. Rev., 2024, doi: 10.1007/s10462-023-10651-9.

[13] L. Qian, J. Guo, H. Qiu, C. Zheng, and L. Ren, "Exploring destination image of dark tourism via analyzing user generated photos: A deep learning approach," Tour. Manag. Perspect., 2023, doi: 10.1016/j.tmp.2023.101147.

[14] X. Cheng et al., "A novel ChatGPT-based multimodel framework for tourism review mining: a case study on China's five sacred mountains," J. Hosp. Tour. Technol., 2024, doi: 10.1108/jhtt-06-2023-0170.

[15] M. Paolanti et al., "Tourism destination management using sentiment analysis and geo-location information: a deep learning approach," Inf. Technol. Tour., vol. 23, no. 2, pp. 241–264, 2021, doi: 10.1007/s40558-021-00196-4.

[16] S. A. Sazan, M. Ahmed, T. B. Saad, and M. Roy, "Advanced Natural Language Processing Techniques for Efficient Sentiment An alysis of US Airline Twitter Data: A High-Performance Framework for Extracting Insights from Tweets," 2024, pp. 1–6. doi: 10.1109/iceeict62016.2024.10534511.

[17] M. Dragoni, M. Federici, and A. Rexha, "ReUS : a Real-time Unsupervised System For Monitoring Opinion Streams," Cognit. Comput., vol. 11, no. 4, pp. 469–488, 2019, doi: 10.1007/S12559-019-9625-X.

[18] S. Kumi, C. C. Snow, R. K. Lomotey, and R. Deters, "Uncovering Concerns of Citizens Through Machine Learning and Social Network Sentiment Analysis," IEEE Access, p. 1, 2024, doi: 10.1109/access.2024.3426329.

[19] Daryanto, I. S. Windiarti, and B. S. Rintyarna, "Social Implications of Technological Advancements in Sentiment Analysis: A Literature Review on Potential and Consequences over the Next 20 Years," Engineering Proceedings, vol. 84, no. 1. p. 49, 2025. doi: 10.3390/engproc2025084049.

[20] F. Z. Xing, B. Schuller, I. Chaturvedi, E. Cambria, and A. Hussain, "Guest Editorial Neurosymbolic AI for Sentiment Analysis," IEEE Trans. Affect. Comput., 2023, doi: 10.1109/taffc.2023.3310856.

[21] Y. Hou, X. Zhong, H. Cao, Z. Zhu, and Y. Zhou, "A shared-private sentiment analysis approach based on cross-modal information interaction," Pattern Recognit. Lett., 2024, doi: 10.1016/j.patrec.2024.05.009.

[22] Z. Liu, B. Zhou, D. Chu, Y. Sun, and L. Meng, “Modality translation-based multimodal sentiment analysis under uncertain missing modalities,” Inf. Fusion, vol. 101, p. 101973, 2023, doi: 10.1016/j.inffus.2023.101973.

[23] C. Zhao et al., “A Systematic Review of Cross-Lingual Sentiment Analysis: Tasks, Strategies, and Prospects,” ACM Comput. Surv., 2024, doi: 10.1145/3645106.

[24] A. H. Ombabi, W. Ouarda, and A. M. Alimi, “Improving Arabic sentiment analysis across context-aware attention deep model based on natural language processing,” Lang. Resour. Eval., 2024, doi: 10.1007/s10579-024-09741-z.

[25] Vetriselvi, “Sentiment analysis on a low-resource language dataset using multimodal representation learning and cross-lingual transfer learning,” Appl. Soft Comput., vol. 157, p. 111553, 2024, doi: 10.1016/j.asoc.2024.111553.

[26] Y. Zhang et al., “A Predictive Model Based on TripAdvisor Textual Reviews: Early Destination Recommendations for Travel Planning,” Sage Open, vol. 14, no. 2, p. 21582440241246430, 2024, doi: 10.1177/21582440241246434.

[27] L. Y. C. Tan, O.-K. Tan, C. C. Sze, and W. W. Bin Goh, “Emotional Variance Analysis: A new sentiment analysis feature set for Artificial Intelligence and Machine Learning applications,” PLoS One, vol. 18, no. 1, pp. e0274299–e0274299, 2023, doi: 10.1371/journal.pone.0274299.

[28] R. M. Davison et al., “The ethics of using generative AI for qualitative data analysis,” Inf. Syst. J., 2024, doi: 10.1111/isj.12504.

[29] G. Roy, “Travelers’ online review on hotel performance – Analyzing facts with the Theory of Lodging and sentiment analysis,” Int. J. Hosp. Manag., vol. 111, p. 103459, 2023, doi: https://doi.org/10.1016/j.ijhm.2023.103459.

[30] J. Recker and A. Perry, “Data on the Margins - Data from LGBTIQ+ Populations in European Social Science Data Archives,” Data Sci. J., vol. 23, p. 39, 2024, doi: 10.5334/dsj-2024-039.

[31] S. Leavy, E. Siapera, and B. O’Sullivan, “Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race,” 2021, pp. 695–703. doi: 10.1145/3461702.3462598.

[32] J. Deeb-Swihart, A. Endert, and A. Bruckman, “Ethical Tensions in Applications of AI for Addressing Human Trafficking: A Human Rights Perspective,” Proc. ACM human-computer Interact., vol. 6, no. CSCW2, pp. 1–29, 2022, doi: 10.1145/3555186.

[33] S. Lewin, R. Chetty, A. R. Ihdayhid, and G. Dwivedi, “Ethical challenges and opportunities in applying artificial intelligence to cardiovascular medicine,” Can. J. Cardiol., 2024, doi: 10.1016/j.cjca.2024.06.029.

[34] O. Papakyriakopoulos and A. Xiang, “Considerations for Ethical Speech Recognition Datasets,” 2023. doi: 10.1145/3539597.3575793.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 738-746. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3052

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

ENHANCING SUSTAINABILITY THROUGH WASTE HEAT UTILIZATION: A STUDY OF BRAZING FURNACE HEAT RECOVERY

Sundaramahalingam Subramaniam*[1], Manikandan Bairavan Veerayan[2]

[1, 2]Department of Electrical and Electronics Engineering, Mepco Schlenk Engineering College, Sivakasi, Tamilnadu, India

[1]https://orcid.org/0000-0002-4776-7546, [2]https://orcid.org/0000-0002-6215-3729

Email: *sundar.eee@mepcoeng.ac.in, bvmani1@mepcoeng.ac.in

ARTICLE INFO

Article History
Received: December 4, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Waste heat recovery,
Sustainability,
Heat Exchanger.

ABSTRACT

Addressing the growing concerns regarding climate change necessitates the decarbonization of energy sectors globally. Heating is the largest energy end-use, accounting for almost half of the total energy consumption in most countries. By partially substituting the use of primary energy sources with waste heat recovery, climate and environmental goals in the future will be more attainable. Industrial processes, such as brazing, generate significant waste heat, which is often released into the environment. This study investigated the feasibility of recovering and repurposing waste heat from a brazing furnace for drying applications. A case study was conducted at a manufacturing facility in which a custom heat recovery system was integrated with a continuous conveyor brazing furnace. The system captures exhaust gases (400–500°C) via a heat exchanger, transferring thermal energy to a drying chamber. Experimental data revealed that 65–70% of waste heat was recovered, reducing the energy consumption of the drying unit by 30% and lowering annual CO_2 emissions by 120 metric tons. The industry utilizes waste heat from the brazing furnace to dry the parts in the washing machine. Economic analysis indicated a payback period of 18 months. This study demonstrates the technical and economic viability of waste heat recovery (WHR) in industrial settings, promoting energy efficiency and sustainability.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

With the majority of nations committed to limiting global warming to 1.5°C, numerous corporations have established individual decarbonization targets for 2030. In 2025, five years remained to implement various decarbonization strategies. However, industrial entities often experience uncertainty regarding the technical pathway for immediate emissions reduction. Technologies that are net present value (NPV) positive and rapidly implementable, such as various energy efficiency measures, can assist organizations in achieving their decarbonization objectives. The recovery of waste heat presents a potential avenue for effectively reducing emissions. Annually, global consumption exceeds 418 exajoules (EJ) or 116,000 terawatt-hours (TWh) of final energy, primarily through the combustion of fossil fuels and heat generation. A portion of this generated heat is utilized for productive purposes, such as electricity production or driving chemical reactions, but the majority remains unused. This unutilized "waste heat" is discharged into the surrounding environment in the form of exhausts or effluents at varying temperature levels.

The recovery of this waste heat can substantially reduce the consumption of primary fuels and, consequently, emissions. The industrial sector accounted for the highest proportion of energy consumption across India in fiscal year 2025 at 41.16%. The domestic sector followed, ranking second, at 25.77%. Traction and railways constituted one percent of the total electricity consumption in India. This data suggests that enhancing energy efficiency through waste heat recovery models can enable Indian businesses to reduce operational costs, improve energy efficiency, and decrease industrial CO2 emissions. With the increasing trend of rising fuel prices over recent decades and growing concerns regarding global warming, engineering industries have been confronted with the challenge of reducing greenhouse gas emissions and improving the efficiency of their facilities. In this context, the utilization of waste heat recovery systems in industrial processes has emerged as a key area of research for reducing fuel consumption, mitigating harmful emissions, and enhancing production efficiency.

Industrial waste heat refers to the energy generated in industrial processes that is not utilized practically and is dissipated or released into the environment. High-temperature processes such as brazing, employed in the automotive and aerospace industries, generate exhaust gases at 400–600°C, which are typically expelled without treatment. Simultaneously, drying operations in these facilities rely on fossil fuels or electricity, contributing to operational costs and emissions. The recovered waste heat could also facilitate a more economical energy transition. Recovered heat can be utilized for industrial process heating purposes, city grid heating, or converted to electricity (determined by local energy demand patterns as well as prices of fuels and electricity). Sources of waste heat primarily include heat loss transferred through conduction, convection, and radiation from industrial products, equipment, and processes, as well as heat discharged from combustion processes. The heat loss can be categorized into high, medium, and low temperature grades. Waste Heat Recovery (WHR) systems are implemented for each range of waste heat to achieve optimal efficiency.

This study addresses the gap in utilizing brazing furnace waste heat for drying, a rarely explored synergy. A case study of a mid-sized automotive parts manufacturer evaluated the design, integration, and performance of a WHR system. The main goal of this work is to study and examine the waste heat recovery from an Industrial Brazing Furnace and by using the obtained energy for useful purposes (here the drying of parts), thereby reducing energy consumption. This unused waste heat represents a significant opportunity for improving energy efficiency and reducing emissions. By implementing waste heat recovery systems, industrial companies can capture and repurpose this energy for various applications, such as preheating input materials, generating electricity, or providing space heating. The potential benefits of waste heat recovery extend beyond individual companies, as they can contribute to broader sustainability goals and help nations meet their climate commitments.

II. RELATED WORKS

Numerous studies have focused on utilizing waste heat from industrial processes for energy-saving purposes [1]. One investigation explored the combination of power cycles to efficiently recover waste heat, particularly in high-temperature thermal furnace units [2]. Another study analysed the performance of a heat pipe heat exchanger using nanofluids to recover waste heat, and demonstrated increased heat transfer coefficients with higher mass concentrations [3]. Additionally, a high-efficiency flue gas condensing heat exchanger was developed to recover waste heat from petrochemical heating furnaces, demonstrating significant energy-saving efficiencies [4], [5]. Furthermore, investigations have examined the optimization of organic Rankine cycles for electricity production using waste heat from industrial processes, suggesting their potential for improving energy efficiency [6-8]. These findings collectively provide insights into potential methods and technologies for utilizing waste heat from industrial brazing furnaces for drying applications [9].

Waste heat recovery from industrial processes offers significant potential for energy savings and greenhouse gas reduction. Studies have indicated that recovering waste heat could substitute for 9% of total energy used by US industry [10]. Various technologies exist for waste heat recovery, including high-temperature vapor compression heat pumps for drying processes, which can reduce energy demand by 60-80% [11]. A novel waste heat recovery-assisted heat pump dryer demonstrated superior performance compared to a simple heat pump dryer, with higher energy and exergy efficiencies, specific moisture extraction rate, and coefficient of performance [12], [13]. This system also exhibited economic viability with a payback period of approximately three years. Waste heat can be recovered from various sources such as exhaust gases, air streams, and liquids, with high-grade sources being more readily applicable for preheating applications [14]. These advancements in waste heat recovery technologies offer promising solutions for improving the industrial energy efficiency and competitiveness.

The escalating global concern regarding climate change necessitates a fundamental shift towards decarbonizing the energy sectors. Heating, as the largest energy end-use, consumes nearly half of the total energy in most countries [15]. Mitigating this substantial energy demand and achieving future climate and environmental goals requires innovative solutions. One promising approach involves harnessing waste heat, which is a significant untapped energy resource often overlooked in industrial processes [16]. This paper presents a case study that investigates the feasibility and effectiveness of recovering waste heat from a brazing furnace to enhance the drying process of an industrial washing machine within a manufacturing setting. The high-temperature waste heat generated by industrial processes such as brazing presents a substantial opportunity for energy efficiency improvements [17]. By effectively capturing and utilizing otherwise wasted energy, it is possible to significantly reduce reliance on primary energy sources, decrease greenhouse gas emissions, and contribute to sustainable energy development [18], [19].

The integration of this system within a broader industrial context aligns with an increasing focus on energy efficiency pathways for decarbonization [20]. The pursuit of efficient waste heat recovery (WHR) has stimulated the development of a diverse array of technologies, each with its own strengths and limitations. A comprehensive understanding of these technologies is essential to select the most appropriate solution for a given application. Organic Rankine Cycles (ORCs), for instance, represent a mature technology for converting waste heat into electricity [21]. They have demonstrated considerable potential in various industrial sectors, notably electric arc furnaces (EAFs) [22], where significant quantities of waste heat are generated. However, despite their potential, widespread adoption of ORCs remains limited. This limitation is partially attributable to the challenges faced by investors and end-users in accurately assessing and realizing the potential energy savings associated with their implementation [23].

Alternative WHR approaches include the use of heat pipes coupled with thermoelectric generators (TEGs) [24]. This combined system offers a compelling solution for bridging the gap between energy supply and demand, effectively managing the inherent mismatch in the energy distribution frequently encountered in industrial environments [25]. The heat pipes excel at efficient heat transfer, whereas the TEGs directly convert the thermal energy into electricity. Supercritical CO2-based power cycles represent another promising avenue for waste heat to power (WH2P) applications [26]. Their applicability extends across a broad spectrum of energy-intensive industries, including steel and metal production, cement manufacturing, and petrochemicals. The selection of the optimal WHR technology depends on several critical factors, including the temperature and quantity of available waste heat, desired application of the recovered energy, and overall economic viability of the chosen system. The specific characteristics of the waste heat stream from the brazing furnace informs the selection of an optimal heat recovery method for this case study.

The relatively high temperature of the waste heat from the brazing furnace may render technologies such as ORCs or supercritical CO2 cycles more suitable, although the specific design and cost-effectiveness require further investigation. However, the focus on drying in a washing machine suggests that a direct heat application might be more appropriate, potentially utilizing heat exchangers or other direct heat transfer methods. This study aims to demonstrate a practical, real-world application of waste heat recovery (WHR) in an industrial context, quantifying its environmental and economic benefits. The successful implementation of this WHR system offers a valuable model for other industries seeking to minimize their environmental impact and enhance operational efficiency.

III. EXPERIMENTAL SETUP

III.1 FURNACE BRAZING

Furnace brazing is an industrial, high-volume brazing technique in which the entire assembly is heated to the melting point of the braze alloy, which is subsequently introduced into the joint before cooling. The primary advantages of furnace brazing include the ability to join dissimilar materials, enhanced control over tolerances, and mitigation of distortion in the finished component. Brazing is a joining method for metals or ceramics that utilizes a metal with a lower melting point than either of the materials to be joined to create a joint. By definition, only processes employing a metal that melts above 450°C are classified as brazing while processes below this temperature are categorized as soldering. In furnace brazing, the entire assembly is heated to the temperature at which the braze alloy melts and flows into the joint. Upon subsequent cooling, it solidified, and formed a robust joint. Furnace brazing enables the simultaneous creation of numerous joints, often numbering in thousands.

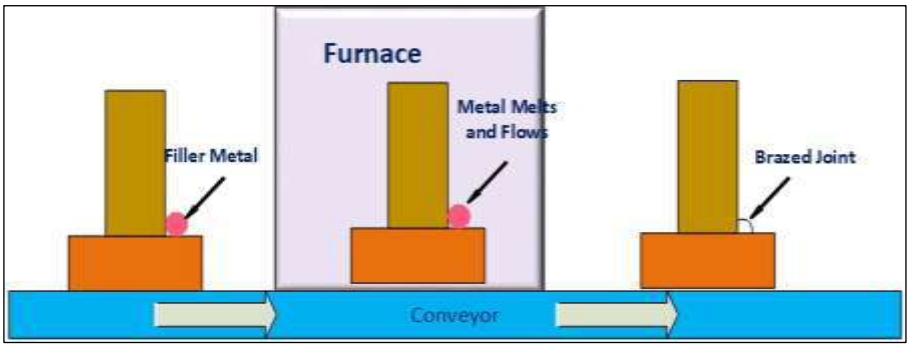


Figure 1: Schematic of furnace brazing process.
Source: Authors, (2026).

Various heating methods are available for brazing operations as shown in Figure 1. The primary consideration in selecting a heating method is to achieve efficient heat transfer throughout the joint while retaining the heat capacity of the individual base metals utilized. The geometry of the braze joint is also a critical factor to consider, as are the rate and volume of production required. A 250 kW natural gas brazing furnace (operating at 850°C, exhaust flow rate 0.8 m³/s) was analysed. The drying chamber required 80°C for drying the polymer component. The detailed functional diagram of brazing furnace is given in Figure 2. The easiest way to categorize brazing methods is to group them using a heating method. Here are some of the most common:

- Torch brazing
- Furnace brazing [Exothermic, Hydrogen, Argon and Vacuum Atmospheres]
- Induction brazing
- Dip brazing.
- Resistance brazing
- Infrared brazing
- Blanket brazing
- Electron beam and laser brazing
- Braze welding

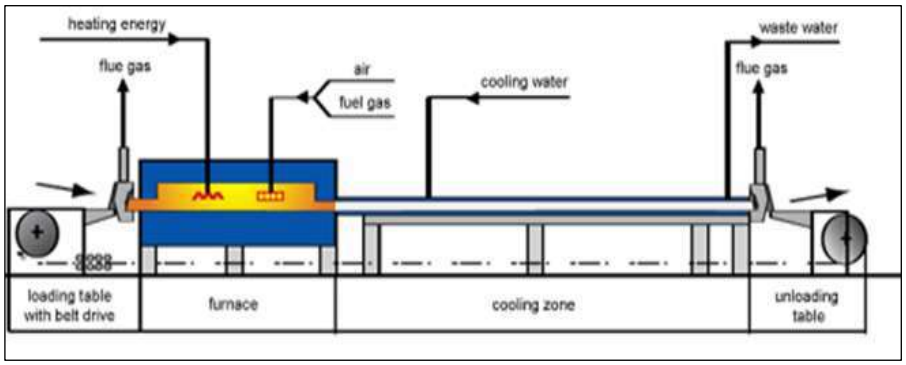


Figure 2: Functional diagram of Brazing Furnace.
Source: Authors, (2026).

III.2 WASHING MACHINE

In the manufacturing industry, washing machines are high-performance inline conveyor washers. The working principle is based on the linear movement of parts through the process; parts are conveyed through different stages of the process at controlled and changeable speed. These washers are suitable for simple geometries with a short quick cycle time per component. The Multistage conveyor parts washer represents a remarkable advancement in industrial cleaning technology, offering an efficient and comprehensive solution for the increasingly complex requirements of component cleaning. This cutting-edge system was designed to meet the needs of diverse industries, from automotive to electronics, and beyond.

The machine uses a series of conveyor belts to transport components through various cleaning stages, each tailored to address specific contamination challenges. The operation of this innovative industrial parts washer machine begins with the loading of components onto the conveyor, which then guides them through a sequence of meticulously engineered cleaning stages. These stages include processes such as pre-washing, degreasing, ultrasonic cleaning, rinsing, drying, and even inspection. At each stage, the components are subjected to specialized cleaning techniques and solutions to ensure that contaminants such as oils, greases, dust, and particulates are thoroughly removed. The modular nature of the multistage system allows for customization, enabling manufacturers to adapt the machine to the unique cleaning requirements of their components.

IV. MATERIALS AND METHODS

The experimental setup included the installation of a heat recovery system in the brazing furnace. The heat exchanger captures the waste heat from the furnace exhaust and channels it into a drying chamber. The setup was designed to ensure an efficient heat transfer while maintaining the operational integrity of the brazing furnace. Based on waste heat characterization, a heat exchanger was designed to efficiently transfer heat from the exhaust gases of the brazing furnace to the water inlet of the industrial washing machine. A shell-and-tube heat exchanger was selected because of its high efficiency and suitability for high-temperature applications. The heat exchanger was designed using commercially available software, considering factors such as the temperature difference between the hot and cold streams, flow rates of the fluids, and overall heat transfer coefficient.

The materials selected for the heat exchanger were based on their resistance to high temperatures and corrosion, considering the chemical composition of the exhaust gases. The heat exchanger was fabricated using high-quality materials and precision manufacturing techniques to ensure durability and performance. The system also includes a control system to regulate the flow of exhaust gases and water, maintain optimal operating temperatures as shown in Figure 3. The control system was programmed to maintain a specific temperature difference between the hot and cold streams, maximizing the heat transfer efficiency while ensuring safe operation.

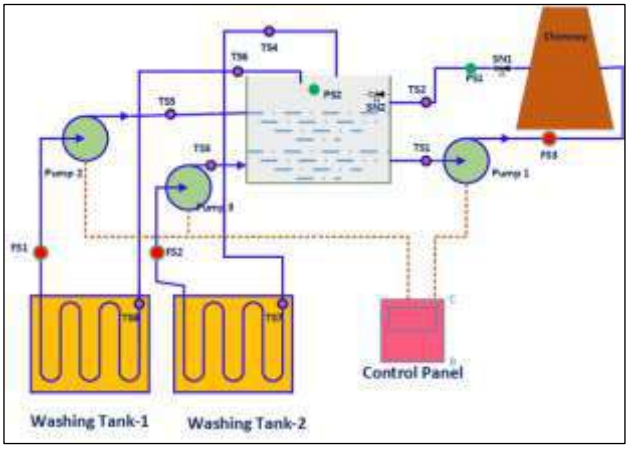


Figure 3: Boundary diagram of the Waste Heat Recovery System.
Source: Authors, (2026).

IV.1 STEP BY STEP PROCEDURE FOR IMPLEMENTATION

IV.1.1 Heat Recovery System Installation

To optimize the efficiency of our brazing furnace, it is essential to install a heat recovery system in the exhaust. The system should be designed to capture and transfer heat from the furnace exhaust to the drying chamber. Insulated ducts must be carefully engineered to ensure the effective transportation of heat. The heat exchanger must be securely attached to the furnace exhaust outlet and meticulously insulated to minimize any potential heat loss. Furthermore, it is imperative to use insulated ducts to connect the heat exchanger to the drying chamber. In addition, the installation of fans or blowers should be considered to facilitate efficient heat transfer and air circulation within the system. Moreover, the design of the drying chamber should be carefully tailored to accommodate specific materials to be dried. It is crucial to ensure that the chamber is well-insulated to effectively retain the captured heat, thereby maximizing the overall efficiency.

IV.1.2 Sensor Installation and Calibration

To ensure precise monitoring of the drying process, it is necessary to install temperature sensors at specific locations. These sensors were placed in the furnace exhaust, before and after the heat exchanger, and inside the drying chamber. Thorough calibration of these sensors is essential to ensure accurate temperature readings. The sensor details and their locations are presented in table 1.

Table 1: Various Sensors and their installation locations.

Temperature sensor Name	Location
TS1	PUMP 1 OUTPUT
TS2	CHIMNEY IN
TS3	PUMP 3 OUTPUT
TS4	WASH TANK 2 OUTPUT
TS5	PUMP 2 OUTPUT
TS6	WASH TANK 1 OUTPUT
TS7	IN WASH TANK 1
TS8	IN WASH TANK 2
TF1	FURNACE OUTLET
TF2	CHIMNEY BODY
Flow sensor Name	Location
FS1	PUMP 2 OUT
FS2	PUMP 3 OUT
FS3	PUMP 1 OUT
Pressure sensor Name	Location
PS1	CHIMNEY IN
PS2	TANK PRESSURE
Solenoid valve Name	Location
SN1	CHIMNEY IN
SN2	TANK IN

Source: Authors, (2026).

Additionally, moisture content analyzers were installed at the entry and exit points of the drying chamber to measure the moisture content of the materials before and after the drying process. The comprehensive data acquisition system is used to connect and integrate all sensors and analyzers. This system will be configured to systematically record data at regular intervals, allowing us to closely monitor and analyze temperature and moisture content variations throughout the entire drying process.

IV.1.3 Data Acquisition and Analysis

The next phase involves collecting data on the performance of the WHR system. Temperature sensors were installed at various points in the system, including the inlet and outlet of the heat exchanger, to monitor the temperature changes of both exhaust gases and water. Flow meters were used to measure the flow rates of both the streams. The energy consumption of the drying system of the washing machine was measured by using a power meter installed on the main power supply. Data were collected over a period of three months, encompassing various operating conditions and loads. Data logging was performed using a data acquisition system capable of recording data at a frequency of 1 Hz. The collected data were then analysed to determine the amount of waste heat recovered, the efficiency of the heat transfer process, and the reduction in energy consumption for the drying cycle of the washing machine.

Statistical analysis techniques were employed to determine the significance of the results and identify the correlations between the various parameters. The data analysis involved calculating the overall heat transfer coefficient, effectiveness of the heat exchanger, and energy savings achieved through waste heat recovery. This comprehensive data analysis provided a robust evaluation of the performance of the WHR system. A convenience sampling method that is nothing but a non-probability sampling technique has been planned, as the samples are selected from the population directly because they are conveniently available for us regularly at any point of a given time. Samples such as the temperature of the furnace, temperature of the cooling water in the washing machine, and level of water is noted based on different criteria such as start-up time, brazing time and full load of the washing machine.

IV.1.4 Baseline Data Collection

This detailed analysis evaluates the effectiveness of utilizing waste heat from an industrial brazing furnace for drying parts. The primary objectives are to assess energy savings, improve drying efficiency, and analyze overall system performance. Data was collected from a controlled experimental setup to compare baseline conditions (without heat recovery) against the improved system (with heat recovery). Operate the brazing furnace under normal conditions without the heat recovery system and collect baseline data on furnace exhaust temperature and material drying times. Data is collected on key parameters, including:

Exhaust Temperature: Temperature of the furnace exhaust before and after heat recovery.
Drying Chamber Temperature: Temperature inside the drying chamber during the drying process.
Drying Time: Time required to achieve the desired moisture content in materials.
Energy Consumption: Total energy used by the brazing furnace and the drying process.
Nature of Data -Two different types of data were collected in this case study. Qualitative data with functional lists, observations, etc., and quantitative data with power readings, parameters, measurements, etc.

Table 2: Specifications of proposed Waste Heat Recovery System.

Water Tank 1 Capacity	550 Litres
Power Consumption per hour	10 kW
Water Tank 2 Capacity	700 Litre
Power Consumption of water tank 2	12 kW
Electricity rate (approx.)	10 Rs./kWH
Average Working hours	12 hrs for 2 shift
Power required for weekly pull up of 1200 Litres from 30oC to 60oC	42 kWH
Energy required for maintain application per hour	22 kW
Per day energy required to maintain temperature considering 2 shifts	264 kW
Time required for pull up considering 30 kW heat pump	1.4
Heat pump consumption	13.2 kW
Running hours	10.2 hrs
Gross daily consumption including pull up	134.64 kW

Source: Authors, (2026).

IV.2 CASE 1: RUN THE BRAZING FURNACE WITHOUT HEAT RECOVERY

When running the brazing furnace, it was ensured that it operated under normal conditions without activating the heat recovery system. Thoroughly document the baseline data, including exhaust temperature, energy consumption, and drying performance, using conventional methods to ensure accuracy and completeness.

IV.3 CASE 2: WITH HEAT RECOVERY OPERATION

To activate the Heat Recovery System, start the heat exchanger and ensure that heat is effectively transferred to the drying chamber. A control system was utilized to maintain the desired temperature levels within the drying chamber, making precise adjustments necessary to create an optimal drying environment. When loading the drying chamber, care should be taken to place the materials to be dried inside the chamber, ensuring that they are evenly distributed to promote uniform drying and the efficient use of space. During the drying process, it is important to continuously monitor temperature, moisture content, and energy consumption. The air flow and temperature settings were adjusted to optimize the drying process, ensuring that the materials were dried effectively while minimizing energy usage. Regular monitoring and adjustments helps to achieve the best possible drying results. Air flow and temperature settings were required to optimize the drying process.

IV.4. GROSS ENERGY CALCULATIONS

To calculate the daily power consumption of all the loads with an hourly time step, a good understanding of how often the loads are used during the two different periods of the day is required. Depending on the specific applications, the battery storage bank is calculated based on battery specifications such as the battery storage capacity (the minimum batteries) for off-grid applications, batteries must fulfil the discharging rate that has to be larger than or equal to the peak power of the load capacity, the battery storage capacity must large enough to supply the longest day time energy used, and to be able to use energy during the longest cloudy period. The gross energy requirement (EG) of an energy conversion device with the energy output (EO), can be decomposed in two parts the direct input of energy during operation (ED), and indirect energy requirement (EI), the energy used by the loads can be calculated by the Equation 1,

$$E_G = E_D + E_I \quad (1)$$

Where, E_G: Gross energy requirement, E_D: Direct input of energy, E_I: Indirect input of energy.

Solar panels and batteries can be calculated using basic equations that depend primarily on the total daily load capacity. One principle is that the solar panels charge the batteries during the day and the batteries process the load at night as in the solar street lighting system. The capacity of storage batteries can be calculated using the following Equation 2,

$$C_B(Ah) = \frac{E_{load}}{\eta_{inv} \, x \, V_B \, x \, DOD} \, x \, DOA \quad (2)$$

Where, E_{load} :Total load Energy (Wh/day), η_{inv} : Inverter efficiency, V_B: Battery voltage (V), DOD: Depth of discharge (%), DOA: Day of autonomy

V. RESULTS AND DISCUSSIONS

The heat exchanger effectively reduced the exhaust temperature from 55 to 250°C, capturing significant waste heat. This indicates an efficient heat transfer and minimal losses, which are vital to the performance. The temperature of the chamber was effectively maintained within the desired range (80-100°C) by using the recovered heat to ensure optimal drying conditions. Implementing the heat recovery system reduced the drying time from 120 minutes to 90 minutes, a 25% improvement. This suggests that a higher temperature in the drying chamber enhanced the evaporation rate of moisture from the parts. Both baseline and heat recovery setups achieved a target final moisture content of 5%, indicating that the heat recovery system maintained or improved the drying effectiveness. Using the heat recovery system resulted in a 30% reduction in energy consumption, saving 50 kWh per drying cycle.

With an average industrial electricity cost of (Indian Rupees) INR 10 per kWh, this translates to INR 50 per drying cycle. For a facility operating 300 drying cycles per year, the annual savings are approximately INR 15,000.

Table 3: Comparison of performance of furnace with/without WHR.

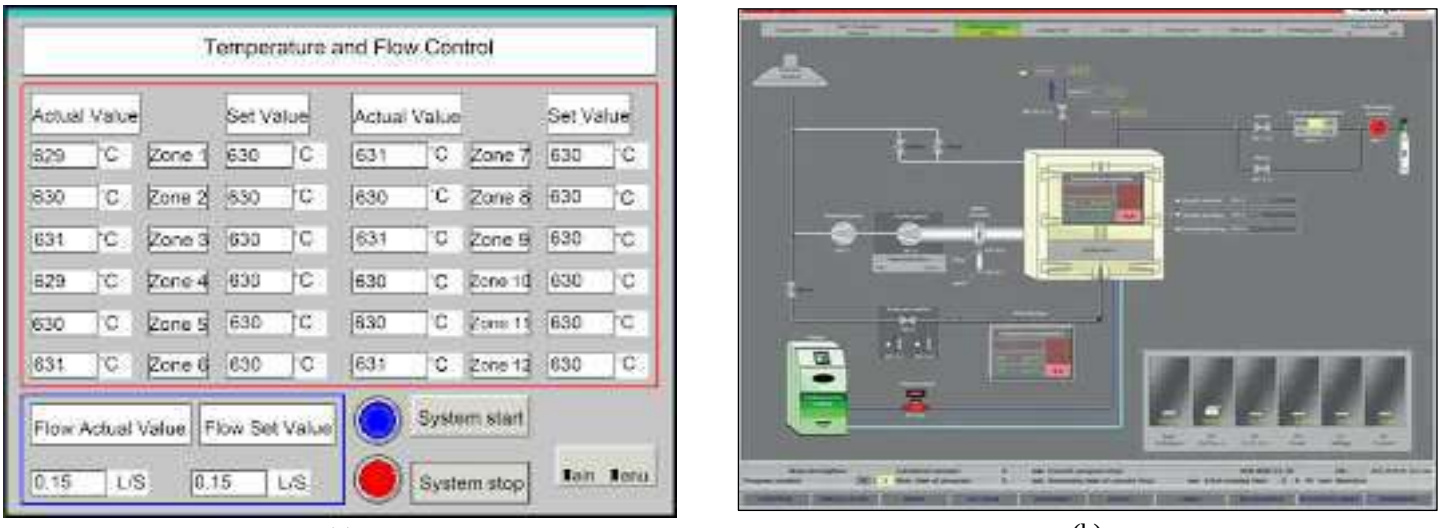
Description	Without Heat Recovery	With Heat Recovery System
Furnace Exhaust Temperature	550°C.	550°C. 250°C after heat exchange
Drying Chamber Temperature	~25°C	80°C and 100°C, optimal for drying
Initial Moisture Content of Parts	20%.	5% after 120 minutes
Final Moisture Content	5% after 120 minutes	5% after 90 minutes
Energy Consumption	150 kWh per drying cycle	100 kWh per drying cycle

Source: Authors, (2026).

Through meticulous analysis, it becomes apparent that harnessing waste heat from the brazing furnace serves as a cornerstone for elevating the energy efficiency quotient of the drying process, concurrently causing substantial reductions in operational expenditures. The system's efficacy manifests prominently through a notable 33% decline in overall energy consumption metrics, emblematic of its adept utilization of surplus thermal energy emanating from the brazing furnace. This profound reduction in energy demand is orchestrated through the sophisticated interplay of heat exchange mechanisms and optimized heat transfer modalities, effectively channeling waste heat streams to augment the thermal energy budget allocated for the drying process. Consequently, this judicious exploitation of latent thermal resources obviates the need for additional energy inputs, thereby mitigating resource wastage and affording tangible economic savings. Moreover, the system's prowess is further accentuated by a consequential 25% reduction in drying duration, indicating of its advantages in expediting the moisture evaporation kinetics intrinsic to the drying process.

This temporal optimization is underpinned by a synergistic amalgamation of enhanced heat transfer dynamics facilitated by the waste heat recovery infrastructure and the meticulous control of the process parameters governing airflow rates, temperature gradients, and moisture content within the drying chamber. Temperature are monitored though control panel as shown in Figure 4 (a) and (b). In summary, the comprehensive analysis underscores the transformative impact of waste heat utilization from the brazing furnace, accentuating its pivotal role in recalibrating the energy efficiency paradigm of the drying process. By delivering tangible reductions in energy consumption and drying duration, the system epitomizes a paradigm shift towards sustainable and economically prudent industrial operations, underscoring its indispensable contribution to operational excellence and resource conservation endeavors.

The inherent scalability of the waste heat recovery system empowers its seamless adaptation and expansion to accommodate a diverse array of industrial processes characterized by elevated temperature regimes and demanding drying prerequisites. By leveraging modular design principles and flexible configuration options, the system can be tailored and customized to align with the specific operational requirements and production dynamics of various industrial applications. Regardless of deployment in metalworking, chemical processing, or food production facilities, the system exhibits unparalleled versatility and adaptability, underscoring its suitability for deployment across a spectrum of high-temperature industrial processes. This scalability not only enhances operational agility and responsiveness but also future-proofs investments by facilitating incremental expansions and upgrades in tandem with evolving production needs and technological advancements.



(a) (b)

Figure 4: Temperature and Flow control from control panel.
Source: Authors, (2026).

During daytime operating loads, solar energy is directly used through the inverter to supply the load without passing through the storage battery. Therefore, the need for a battery depends on the instantaneous level of energy generated from the solar panels. When the generated energy is less than the load requirement, the battery compensates for the energy shortage, and when the energy generated is higher than the load requirements, the battery is charged.

VI. CONCLUSIONS AND RECOMMENDATIONS

In summary, this study of recovering waste heat from industrial brazing furnaces for drying applications underscores the paramount importance of optimizing the thermal resources in industrial processes to enhance energy efficiency, decrease operational expenses, and reduce environmental harm. The investigation reveals the substantial benefits of employing innovative technologies such as advanced heat exchangers, thermoelectric generators, hybrid systems, and Industry 4.0 tools. These technologies have proven effective in maximizing energy use and minimizing waste, with experimental outcomes showing a recovery of 65-70% of waste heat, leading to a 30% reduction in energy consumption of the drying unit and a notable decrease in CO_2 emissions by 120 metric tons annually.

The research provides a comprehensive overview of the current advancements in waste heat recovery and lays down a strategic pathway for further exploration and technological development in this field. To propel waste heat recovery forward, the study recommends sustained investment in research and development, the adoption and integration of new technologies, comprehensive training and education for industry professionals, enhanced regulatory support and incentives, continuous system monitoring and optimization, and increased knowledge sharing and collaboration among stakeholders. By adhering to these strategies and adopting an integrated approach to waste heat recovery, industrial entities can achieve considerable energy savings, mitigate environmental impacts, and bolster their competitiveness in the journey towards a more sustainable industrial landscape.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Sundaramahalingam Subramaniam, and Manikandan Bairavan Veerayan.
Methodology: Sundaramahalingam Subramaniam and Manikandan Bairavan Veerayan..
Investigation: Sundaramahalingam Subramaniam and Manikandan Bairavan Veerayan..
Discussion of results: Manikandan Bairavan Veerayan.
Writing – Original Draft: Sundaramahalingam Subramaniam.
Writing – Review and Editing: Manikandan Bairavan Veerayan..
Resources: Manikandan Bairavan Veerayan..
Supervision: Manikandan Bairavan Veerayan..
Approval of the final text: Manikandan Bairavan Veerayan.

VIII. ACKNOWLEDGMENTS

Authors would like to express their thankful to Mepco Schlenk Engineering College, Sivakasi, India for providing necessary facilities to carry out this research

IX. REFERENCES

[1] Y. Lan, X. Zhao, W. Zhang, L. Mu, and S. Wang, "Investigation of the waste heat recovery and pollutant emission reduction potential in graphitization furnace," Energy, vol. 245, p. 123292, Jan. 2022, doi: 10.1016/j.energy.2022.123292.

[2] A. Serafino, P. Cinnella, and B. Obert, "Multi-fidelity robust design optimization of an ORC turbine for high temperature waste heat recovery," Energy, vol. 269, p. 126538, Jan. 2023, doi: 10.1016/j.energy.2022.126538.

[3] M. Elbakhshwan et al., "Advancing Diffusion Bonded Compact Heat Exchangers for High Temperature Applications," AM&P Technical Articles, vol. 181, no. 4, pp. 14–18, May 2023, doi: 10.31399/asm.amp.2023-04.p014.

[4] K. Abrosimov, A. Baccioli, and A. Bischi, "Extensive techno-economic assessment of combined inverted Brayton – Organic Rankine cycle for high-temperature waste heat recovery," Energy, vol. 211, p. 118406, Aug. 2020, doi: 10.1016/j.energy.2020.118406.

[5] A. M. Pantaleo, J. Fordham, O. A. Oyewunmi, P. De Palma, and C. N. Markides, "Integrating cogeneration and intermittent waste-heat recovery in food processing: Microturbines vs. ORC systems in the coffee roasting industry," Applied Energy, vol. 225, pp. 782–796, May 2018, doi: 10.1016/j.apenergy.2018.04.097.

[6] R. R. Del Mundo, "Chapter 223 - A Hybrid Solar-Biomass Drying System for Ceramic Wares," in World Renewable Energy Congress VI, elsevier, 2000, pp. 1091–1093. doi: 10.1016/b978-008043865-8/50223-3.

[7] U. Umoga et al., "Exploring the potential of AI-driven optimization in enhancing network performance and efficiency," Magna Scientia Advanced Research and Reviews, vol. 10, no. 1, pp. 368–378, Feb. 2024, doi: 10.30574/msarr.2024.10.1.0028.

[8] D.-X. Li, S.-S. Zhang, and G.-H. Wang, "Selection of organic Rankine cycle working fluids in the low-temperature waste heat utilization," Journal of Hydrodynamics, vol. 27, no. 3, pp. 458–464, Jun. 2015, doi: 10.1016/s1001-6058(15)60504-2.

[9] T. G. G. Uthpala, S. B. Navaratne, and A. Thibbotuwawa, "Review on low-temperature heat pump drying applications in food industry: Cooling with dehumidification drying method," Journal of Food Process Engineering, vol. 43, no. 10, Aug. 2020, doi: 10.1111/jfpe.13502.

[10] H. Yu, T. Gundersen, and X. Feng, "Process integration of organic Rankine cycle (ORC) and heat pump for low temperature waste heat recovery," Energy, vol. 160, pp. 330–340, Jul. 2018, doi: 10.1016/j.energy.2018.07.028.

[11] Q. Pan, J. Peng, and R. Wang, "Experimental study of an adsorption chiller for extra low temperature waste heat utilization," Applied Thermal Engineering, vol. 163, p. 114341, Sep. 2019, doi: 10.1016/j.applthermaleng.2019.114341.

[12] K. Klotsche, F. Micus, U. Hesse, and C. Thomas, "Waste Heat Recovery for Reciprocating Compressors," IOP Conference Series: Materials Science and Engineering, vol. 604, no. 1, p. 012085, Aug. 2019, doi: 10.1088/1757-899x/604/1/012085.

[13] J. Salazar, "21 - Life cycle assessment (LCA) of windows and window materials," in Eco-efficient construction and building materials, elsevier, 2014, pp. 502–527. doi: 10.1533/9780857097729.3.502.

[14] M. O. Agho, N. L. Eyo-Udo, E. C. Onukwulu, A. K. Sule, and C. Azubuike, “Digital Twin Technology for Real-Time Monitoring of Energy Supply Chains,” International Journal of Research and Innovation in Applied Science, vol. 9, no. 12, pp. 564–592, Jan. 2025, doi: 10.51584/ijrias.2024.912049.

[15] J.-N. Gerdes, M. Munder, and A. Sauer, “Evaluation of technical and economic potential of waste heat distribution networks in industrial sites,” Energy Reports, vol. 9, pp. 219–226, Jan. 2023, doi: 10.1016/j.egyr.2022.12.112.

[16] G. Legorburu and A. D. Smith, “Energy modeling framework for optimizing heat recovery in a seasonal food processing facility,” Applied Energy, vol. 229, pp. 151–162, Aug. 2018, doi: 10.1016/j.apenergy.2018.07.097.

[17] Y. Chen, A. Ozaki, A. Yusuke, Y. Choi, and H. Lee, “Numerical modeling of laundry drying in a commercial laundry plant: Incorporating water potential as a driving force for heat and mass transfer,” Journal of Cleaner Production, vol. 423, p. 138735, Sep. 2023, doi: 10.1016/j.jclepro.2023.138735.

[18] W. Chen et al., “Carbon-neutral heat supply strategies for industrial decarbonization: A critical and systematic literature review,” Cleaner and Responsible Consumption, vol. 15, p. 100244, Dec. 2024, doi: 10.1016/j.clrc.2024.100244.

[19] A. M. Maghrabi, J. Song, and C. N. Markides, “How can industrial heat decarbonisation be accelerated through energy efficiency?,” Applied Thermal Engineering, vol. 233, p. 121092, Jun. 2023, doi: 10.1016/j.applthermaleng.2023.121092.

[20] D. Dadi, M. Benedetti, and V. Introna, “Decarbonization of Heat through Low-Temperature Waste Heat Recovery: Proposal of a Tool for the Preliminary Evaluation of Technologies in the Industrial Sector,” Sustainability, vol. 14, no. 19, p. 12626, Oct. 2022, doi: 10.3390/su141912626.

[21] M. Ngcukayitobi, L. K. Tartibu, and F. Bannwart, “Enhancing Thermo-Acoustic Waste Heat Recovery through Machine Learning: A Comparative Analysis of Artificial Neural Network–Particle Swarm Optimization, Adaptive Neuro Fuzzy Inference System, and Artificial Neural Network Models,” AI, vol. 5, no. 1, pp. 237–258, Jan. 2024, doi: 10.3390/ai5010013.

[22] M. Andayesh, R. Germanier, F. Maréchal, M. Gatti, and D. A. Flórez-Orrego, “Improved Waste Heat Management and Energy Integration in an Aluminum Annealing Continuous Furnace Using a Machine Learning Approach.,” Entropy, vol. 25, no. 11, p. 1486, Oct. 2023, doi: 10.3390/e25111486.

[23] A. I. Turja et al., “Machine learning-based multi-objective optimization and thermal assessment of supercritical CO2 Rankine cycles for gas turbine waste heat recovery,” Energy and AI, vol. 16, p. 100372, Apr. 2024, doi: 10.1016/j.egyai.2024.100372

[24] F. Guo, X. Zhu, P. Li, and X. Yang, “Low-grade industrial waste heat utilization in urban district heating: Simulation-based performance assessment of a seasonal thermal energy storage system,” Energy, vol. 239, p. 122345, Oct. 2021, doi: 10.1016/j.energy.2021.122345

[25] A. Inayat, “Current progress of process integration for waste heat recovery in steel and iron industries,” Fuel, vol. 338, p. 127237, Jan. 2023, doi: 10.1016/j.fuel.2022.127237.

[26] H. Jouhara, N. Khordehgah, S. Almahmoud, B. Delpech, A. Chauhan, and S. A. Tassou, “Waste heat recovery technologies and applications,” Thermal Science and Engineering Progress, vol. 6, pp. 268–289, Apr. 2018, doi: 10.1016/j.tsep.2018.04.017.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 747-762. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3055

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

TRENDS, COLLABORATION NETWORKS, AND RESEARCH IMPACT IN THERMAL MANAGEMENT OF ELECTRONIC DEVICES USING PASSIVE COOLING TECHNIQUE

Nilesh Gaikwad[1], Narendra Deore[2] and Mahadev Madgule[3]

[1, 2, 3]Pimpri Chinchwad College of Engineering: Pune, Maharashtra, India

[1]https://orcid.org/0000-0002-4087-3266, [2]https://orcid.org/0000-0003-4710-3798, [3]https://orcid.org/0000-0001-6584-3787

Email: nilesh.gaikwad@pccoepune.org, narendra.deore@pccoepune.org, mahadev.madgule@pccoepune.org

ARTICLE INFO

Article History
Received: December 5, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Thermal management of Electronic Devices (TMED),
Two-phase cooling,
Vapour chamber,
Passive cooling,
Bibliometric review

ABSTRACT

Thermal management of electronic devices (TMED) has gained attention the worldwide trend of automating processes employing artificial intelligence approaches. Unless efficient thermal management improvements and products that can withstand increased heat loads in small areas are introduced to the market, this field will continue to advance. This study's goal is to provide a bibliometric analysis of passive cooling methods for electronic device thermal management using Scopus publishing records from the previous 25 years. a combination of best-performing nations, content and language types, author collaboration and citations, the study examines the article. Authors cooperation and concurrences, as well as keyword-based and country-wise network analysis, can be observed using the VoS viewer program. The analysis clearly shows that China has the highest contribution (32%) in terms of publications, followed by the United States of America. The Scopus-indexed documents increased by 22 times during the period of 2000 to 2024 and a total of 616 documents received citations of 16563 in the year 2024 in the domain of TMED, which shows the growing interest of researchers. Future scholars will have a thorough understanding of how heat management in electronic equipment has changed over the previous 20 years thanks to the assembled study.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The famous "Moore's Law" was proposed in the 1960s as the integrated circuit industry expanded quickly[1]. Although "Moore's Law" is essentially an empirical relationship in manufacturing instead of a natural rule, its correctness has been proven for decades, and the semiconductor sector regularly uses the law to guide research and development goals. The integrated circuit's capacity to support transistors will continue to grow following "Moore's law" [2], [3]. While driving computational power, the exponential growth in transistor density simultaneously intensifies the thermal design power per unit area, posing a challenge to conventional cooling methods.

Heat dissipation presents a significant challenge in the current digital revolution, as electronic products, particularly smartphones, tablet computers, and data centres, increasingly evolve towards higher performance, greater integration, and miniaturisation [4], [5].This escalation in heat flux demands innovative solutions beyond traditional thermal management, propelling the industry towards advanced three-dimensional packaging and novel material science applications [6]. The stringent cooling requirements for high-integration power electronics systems necessitate significant energy consumption and have recently raised concerns regarding carbon footprints [7]. Notice that almost 40% of energy consumption for the current data centres [8] was used in their cooling infrastructure, which consumes almost 12% of total electricity [9].

The continuous improvement in integrated circuit manufacturing processes has led to increased heat generation, necessitating the development of efficient thermal management technologies to enhance performance and reliability [10]. This growing thermal burden necessitates a paradigm shift from traditional cooling to advanced thermal management solutions, leveraging innovative

materials and architectural designs to effectively dissipate the concentrated heat loads [11]. Passive cooling techniques, like natural convection, heat pipes, vapour chambers, and phase-changing materials, are being increasingly explored to address the thermal challenges posed by compact electronic devices [12].

The use of phase change material and heat pipes offers high effective thermal conductivity [13] and is particularly attractive due to their lack of active power consumption and mechanical complexity, which enhances system reliability and reduces operational expenses. Heat pipes' capacity to passively transmit heat from the evaporating end to the condensing end by using the latent heat of the working fluid has drawn constant focus as a highly effective and reliable two-phase heat transfer technique. [14], [15]. This property makes them suitable for applications where minimal noise, vibration, and maintenance are critical design considerations [16]. Figure 1 shows the active and passive cooling technique for electronic cooling applications.

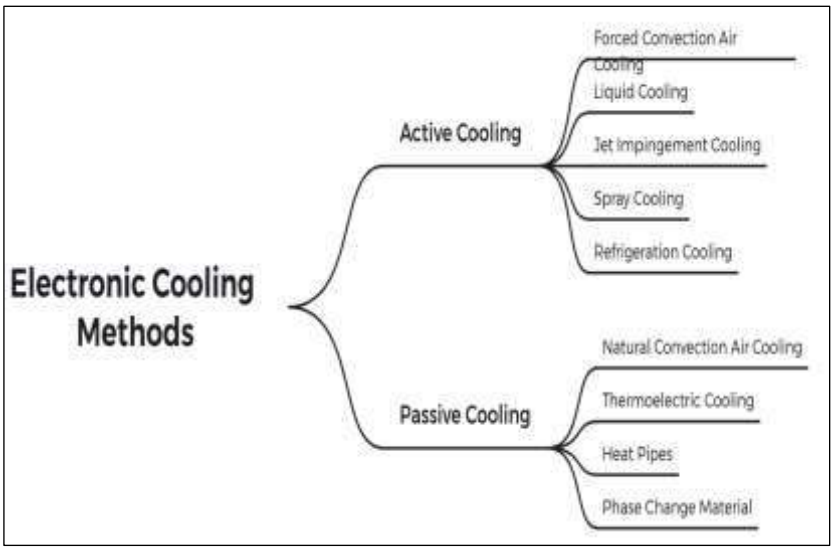


Figure 1: Different techniques for electronic cooling.
Source: [5].

The flat-plate heat pipe, the thermal ground plane, and the heat spreader are synonyms for the vapour chamber, which is two-dimensional and spreads heat from smaller heater areas to larger sink areas across small temperature differences [17]. The fundamental difference between a heat pipe and a vapour chamber is that the former acts as a heat transporter, and the latter serves the purpose of spreading the heat [18]. As shown in Figure 2, vapour chambers (VC), or flat heat pipes, are sealed containers made from materials such as aluminium or copper. The bottom surface is called an evaporator, which is placed on the heat source/chip to be cooled. The top surface is known as a condenser, wherein heat is dissipated to the atmosphere via a cooling medium like air or water. The evaporator and condenser surfaces are internally covered with a thin layer of wick of different types, such as grooves, screen mesh, or sintered powder layer. Non-condensable gases are removed using a vacuum pump, which is then charged with working fluid. The working fluid evaporates due to heat from the source. Vapours travel to the condenser side due to the difference in vapour pressure, and they lose latent condensation heat on the condenser surface. The condensate is transported back to the evaporator by capillary action through the wick.

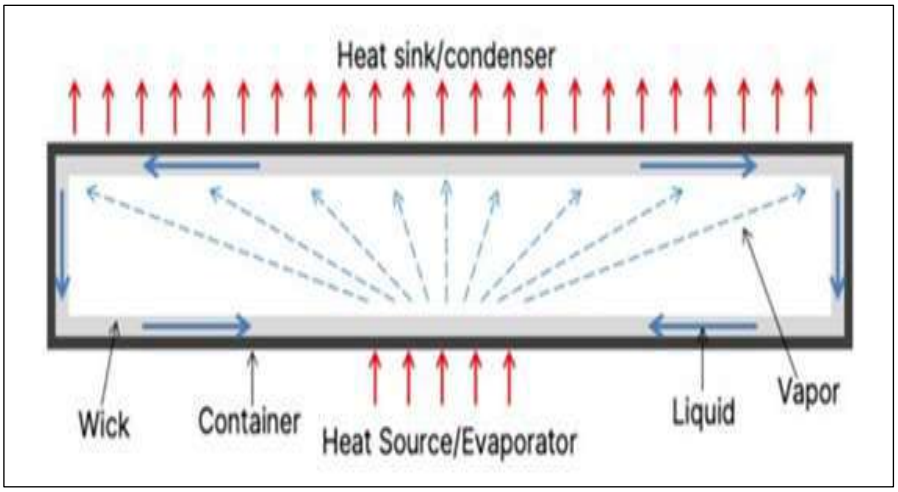


Figure 2: Working of the vapour chamber as a heat spreader.
Source: Authors, (2026).

The excellent thermal dissipation performance of ultra-thin heat pipes (UTHPs) makes them an attractive solution. These advantages make UTHPs widely applicable in industries such as battery thermal management [19], solar collectors [20], [21], and electronics cooling [22].

I.1 CHALLENGES IN USING PASSIVE COOLING TECHNIQUES:

Passive cooling technologies serve as essential heat management solutions for contemporary electronics and industrial applications. Heat pipes, phase change materials (PCMs), and vapour chambers present compelling substitutes for active cooling systems owing to their energy reliability and effectiveness [23]. Nonetheless, despite their theoretical benefits, these technologies encounter significant obstacles that hinder their actual application and market feasibility. The operational limit of a heat pipe indicates the maximum heat transport capacity it can maintain before performance declines, dictated by various physical constraints [24].

These constraints include the capillary limit, where the wick structure fails to return adequate liquid to the evaporator against gravitational and viscous forces; the sonic limit, which arises when vapour velocity nears sonic speed, resulting in choked flow; the boiling limit, where excessive heat flux induces nucleate boiling in the wick and disrupts liquid flow; and the entrainment or flooding limit, where high vapour shear entrains liquid droplets, obstructing return flow. Each limit is influenced by the heat pipe's working fluid, wick geometry, operating orientation, and temperature range [25].

Phase change material suffers operational challenges like low thermal conductivity [26], expansion/contraction during melting/solidification can cause mechanical stresses, leakage, or damage to encapsulation [27], repeated melt-freeze cycles can cause chemical instability, agglomeration, or loss of latent heat storage capacity [28], matching PCM melting temperature with component safe-operating limits is critical; mismatch can lead to thermal throttling or insufficient protection [29].

For vapour chambers, manufacturing complexity increases with the miniaturisation requirements. The integration of micropillar wick structures requires advanced manufacturing capabilities that may not be readily available [30]. Despite these limitations, the heat transfer coefficient owing to the phase change mechanism in these devices makes them a preferred candidate for TMED [31]. Figure 3 compares the heat transfer coefficient values for free convection, forced convection, and phase change mechanism techniques.

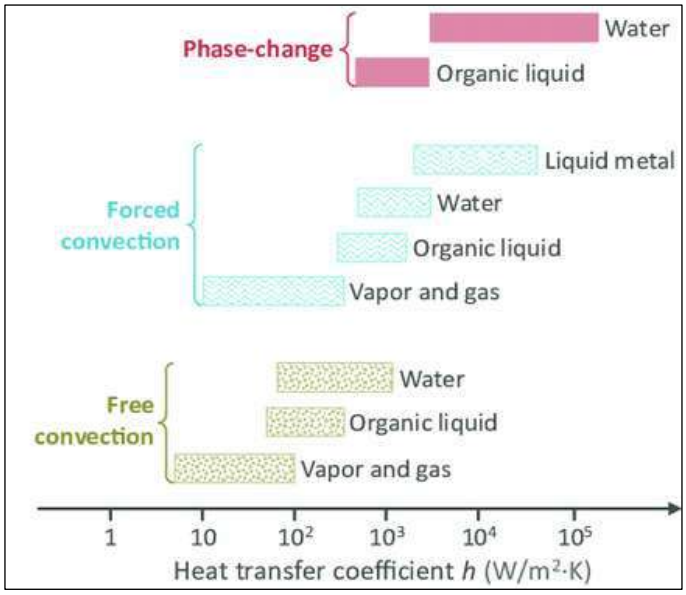


Figure 3: High heat transfer coefficient values for the phase change mechanism.
Source: [32].

The objective of this paper is to present a comprehensive evaluation that includes a bibliometric analysis to gain an understanding of the predominant research efforts and the components that comprise them. A complete analysis of global applications, authors, and published studies is provided in the ensuing sections of the bibliometric analysis. A bibliometric study is defined as a statistical or quantitative analysis of literature, including books, journals, conference proceedings, and online information, to provide insights into a particular subject area or written communication [33]. A bibliometric evaluation offers researchers a robust, data-driven perspective to comprehend the intellectual landscape, discern new trends, drive strategic decisions, and effectively scan extensive literature. When employed efficiently in conjunction with qualitative evaluations, it transforms into an essential instrument in contemporary scholarship. The author found that there is hardly any bibliometric review done in the field of TMED and believes this analysis will help new researchers to understand the trends in publications, author concurrence, and major contributing countries in the area of thermal management of electronic devices.

II. BIBLIOMETRIC SURVEY

An essential step in the research process is gathering data. Scopus, Clarivate, SCImago, Mendeley, ScienceDirect, and Google Scholar are among the most frequently used publishing databases [34]. Scopus is the premier repository of citations and abstracts for peer-reviewed scholarly articles. The Scopus database was consulted to gather literature on Thermal Management of Electronic Devices (TMED), including articles, conference proceedings, and research papers published from 2000 to 2025. In any bibliometric research, choosing the right keywords is essential. The use of keywords often illustrates heterogeneity in the usage of the conjunctions 'AND' and 'OR.' The Scopus database advises that AND should be included when the analysis requires the use of both terms. The conjunction "OR" should be used when one or more words are included in the expected search results [35].

II.1 METHODOLOGY

For the present study shown in Figure 4, Scopus published documents are collected for the last 25 years, and a bibliometric analysis is conducted for the parameters like top ten countries, languages, document types, and citation trends.

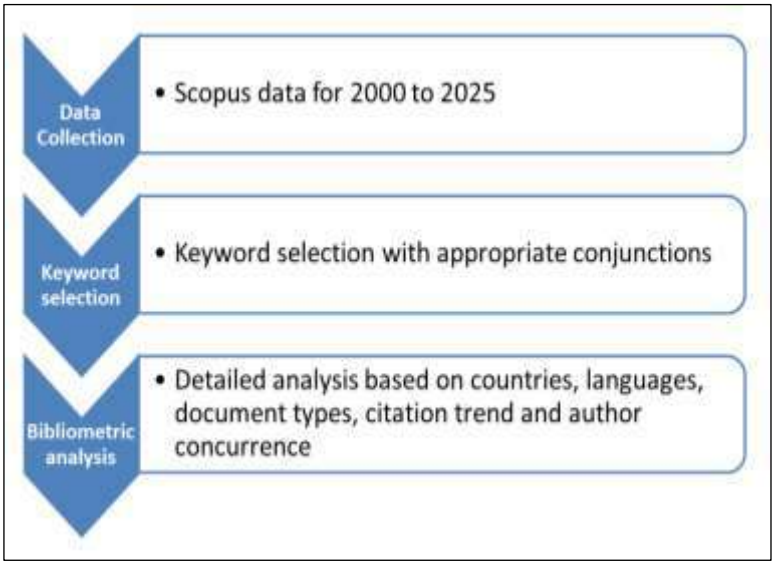


Figure 4: Methodology for the present study.
Source: Authors, (2026).

II.1.1 Data Collection

The bibliometric analysis utilised data sourced from the Scopus website, recognised as a leading and extensive abstract and citation database for peer-reviewed literature, encompassing conference proceedings, book chapters, and articles. Scopus provides a thorough overview of international research output across various disciplines, including science, technology, medicine, social sciences, mathematics, and the arts and humanities, along with advanced tools for tracking, analyzing, and visualizing research. Table 1 presents the keywords utilized for searching Scopus. The bibliometric search keywords were carefully chosen to follow a funnel pattern, starting with broad terms about thermal management and getting more detailed as we looked for new cooling techniques.

This layout lets the frequency analysis track how research interests gradually narrow over time, showing how the field changes from broad ideas to more specific innovations. A total of 3,949 publications were identified from 2000 to 2025 for the search terms Thermal, Management, Electronic, and Devices. A total of 223 papers were identified after limiting the search to TMED with a focus on passive cooling. Upon refining keyword combinations related to TMED utilising passive cooling and vapour chambers, only 16 papers were identified. Scopus was searched with a combination of essential keywords, as shown in Table 1.

Table 1. Bibliometric distribution of technological combinations utilised in TEMD using passive cooling and vapour chamber.

S.N	Search category	Number of **publications**	Year
I.	thermal AND management AND electronic AND devices	3949	2000 to 2025
II.	thermal AND management AND electronic AND devices AND passive AND cooling	223	2000 to 2025
III.	thermal AND management AND electronic AND devices AND passive AND cooling AND Vapour chamber	16	2010 to 2024

Source: Authors, (2026).

The keywords were refined as outlined in Table 1 to examine both the breadth and depth of thermal management mechanisms employing passive cooling techniques and to assess the penetration of vapour chamber technology—a method widely used for efficient heat spreading in compact electronic systems.

II.1.2 Analysis of Bibliometric Data

As discussed in the previous section, the following section analysis is carried out in categories as:

Category -I: ----Thermal Management of Electronic devices;
Category -II: --- Thermal Management of Electronic devices with passive cooling;
Category -III: ---Thermal Management of Electronic devices with passive cooling and vapour chamber.

The study of global research provides insights into nations, languages, kinds of published documents and their publishing patterns, subject-specific evaluations, authors' contributions to research in certain areas, and correlates these with national characteristics. A detailed discussion of the data collected on each of these areas is provided in the sections that follow.

II.1.3 Studies on Thermal Management of Electronic Devices

(a) Research trends in top ten countries

Keyword search using the words “Thermal Management of Electronic devices” returned contributions by 57 different countries in between 2000 to 2025. The search results were plotted for the to 10 top countries. The country-wise distribution of documents eliciting the number of research articles on TMED is shown in Figure 5, displays data on the 10 most productive countries.

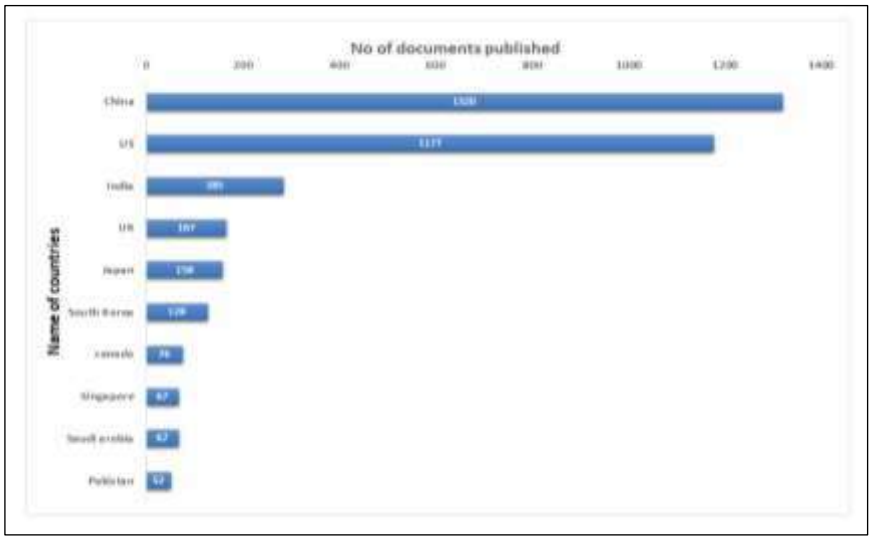


Figure 5: Country-wise research trend from 2000 to 2025.
Source: Authors, (2026).

Figure 5 shows that China leads in terms of publications (1320, 33.43%), followed by the United States of America (1177, 29.80%). The remaining countries are behind in terms of publishing the research in such a crucial domain.

(b) Collaboration between Countries

Data acquired from the Scopus database using VOS Viewer [36] version 1.6.18. Bibliographic coupling research was used to generate network maps linking nations based on their joint publications. The thickness of the lines linking the nodes in each network map is dictated by the "link strength" and signifies network connectedness or co-citations. Fifty-seven countries have been effectively represented on a network analysis diagram. Figure 6 illustrates the network of many nations and their collaborations with coauthors. minimum threshold of five papers was established for a nation to be included in the list.

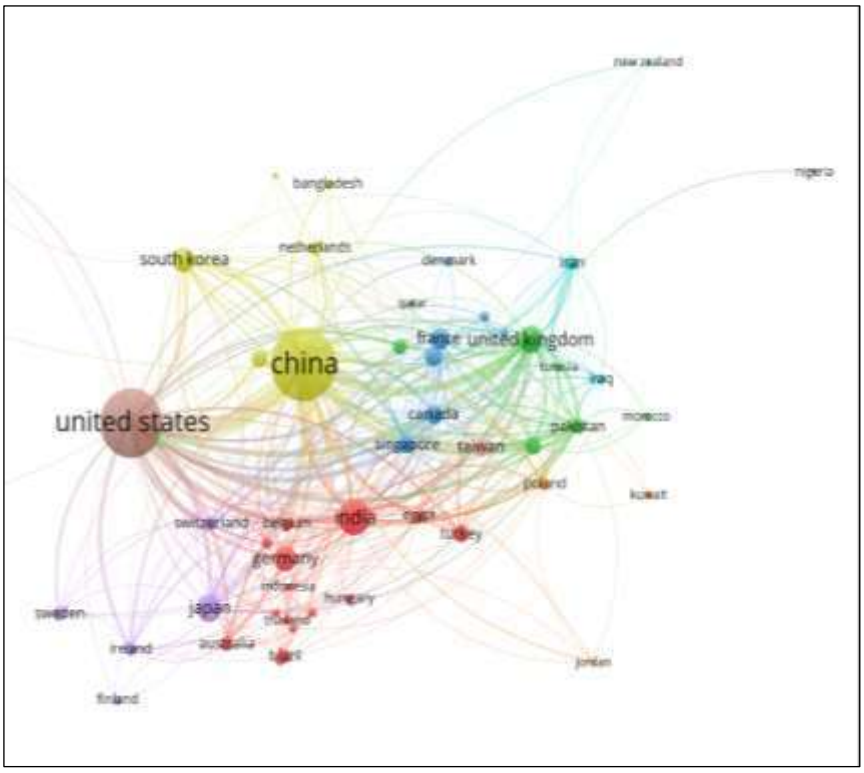


Figure 6: Network analysis diagram based on country keywords.
Source: Authors, (2026).

(c) Languages and Document Types

This section discusses the languages and document types seen in the subject matter. The documents were mainly published in four languages, namely English (3863), Chinese (74), French (5), Japanese (5), Korean (3), Turkish (2), Polish (2) and 1 paper each in Spanish, Italian and German. English accounted for more than 98% of total publications. In terms of document type, book chapters (104), conference reviews (66), short surveys (6), articles (2232), conference papers (1351), reviews (176), and other document types contribute far less to this field of study as shown in Figure 7. The majority of the additional analysis was conducted using document-type articles, which are commonly found in these journals.

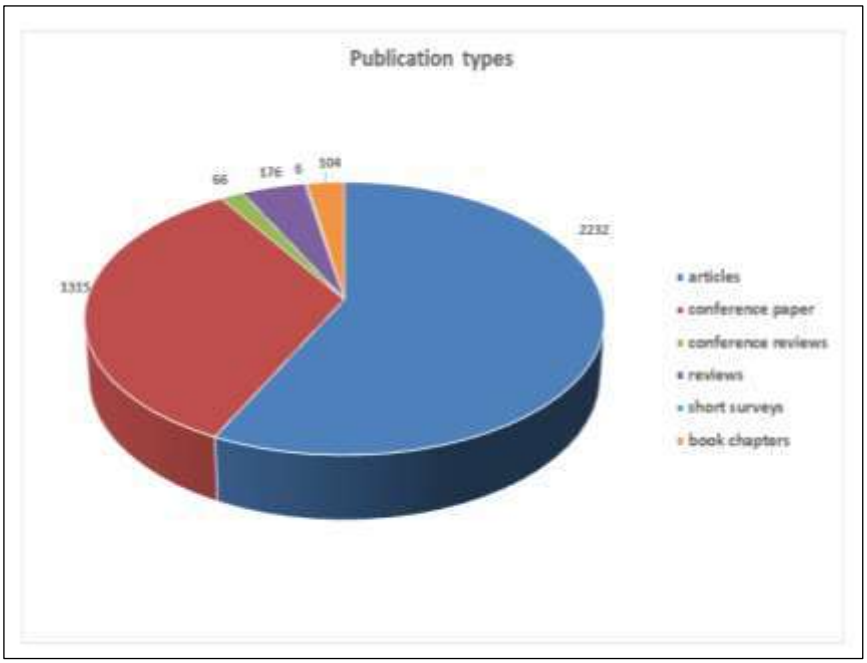


Figure 7: Document type distribution on TMED.
Source: Authors, (2026).

d) Publication Trend

Figure 8 shows the publication trend increasing rapidly post-2015. The year 2024 received a maximum publication of 617 on the topic of TMED. The trendline shows the research initiatives taken by various authors across the globe to address the challenges faced by the electronic community on the thermal front.

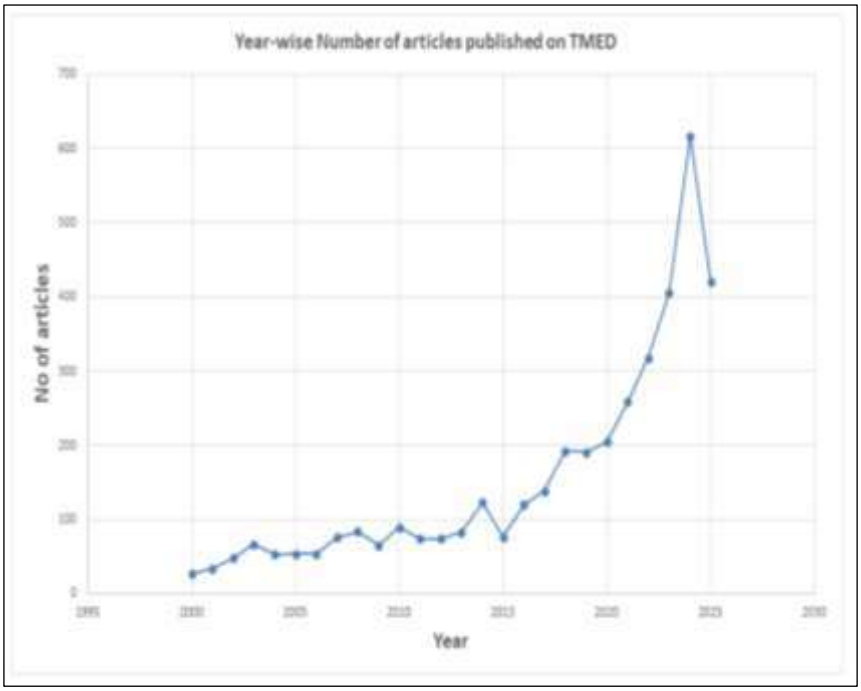


Figure 8: Publication trend year-wise on TMED.
Source: Authors, (2026).

e) Subject area analysis

Though the article's database referred to in this study is limited to “Engineering” subject area only. But similar research is carried out in other subject areas, which is visible from Figure 9 as shown below.

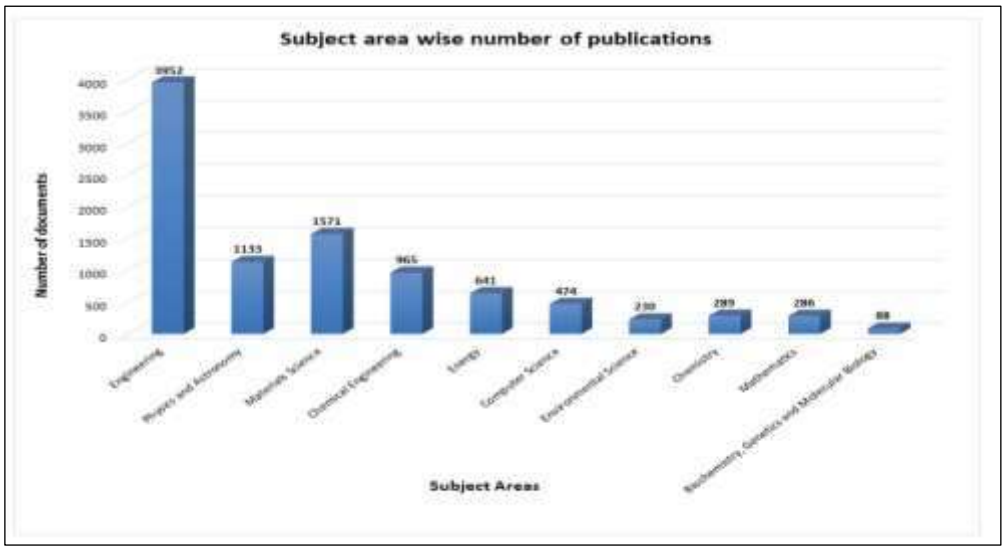


Figure 9: Subject area-wise number of publications on the topic of TMED.
Source: Authors, (2026).

f) Keyword-Based Network Analysis and Author Collaboration

Scopus data was further utilized for network analysis using VOS viewer version 1.6.18. As shown in Figure 10, network analysis for co-occurrences of indexed terms is displayed. When analysing the data, the indexed keywords took into account synonyms, plurals, and variations in spelling. More frequent keywords are denoted by larger font sizes on the network analysis graph, with five being the minimal number of occurrences. The data allowed 2426 out of the total 22803 keywords to be plotted. The term "temperature control" had the highest frequency, which had an effect on other keywords from other clusters that were represented by various colours. The term "electronic cooling" appears frequently and has been connected to other keywords. The terms "thermal conductivity," "heat sinks," "phase change materials," "semiconductors," "electronic packaging," "two phase," and "microelectronics" were significantly prevalent. Thus it appears that various aspects of TMED are explored by various authors.

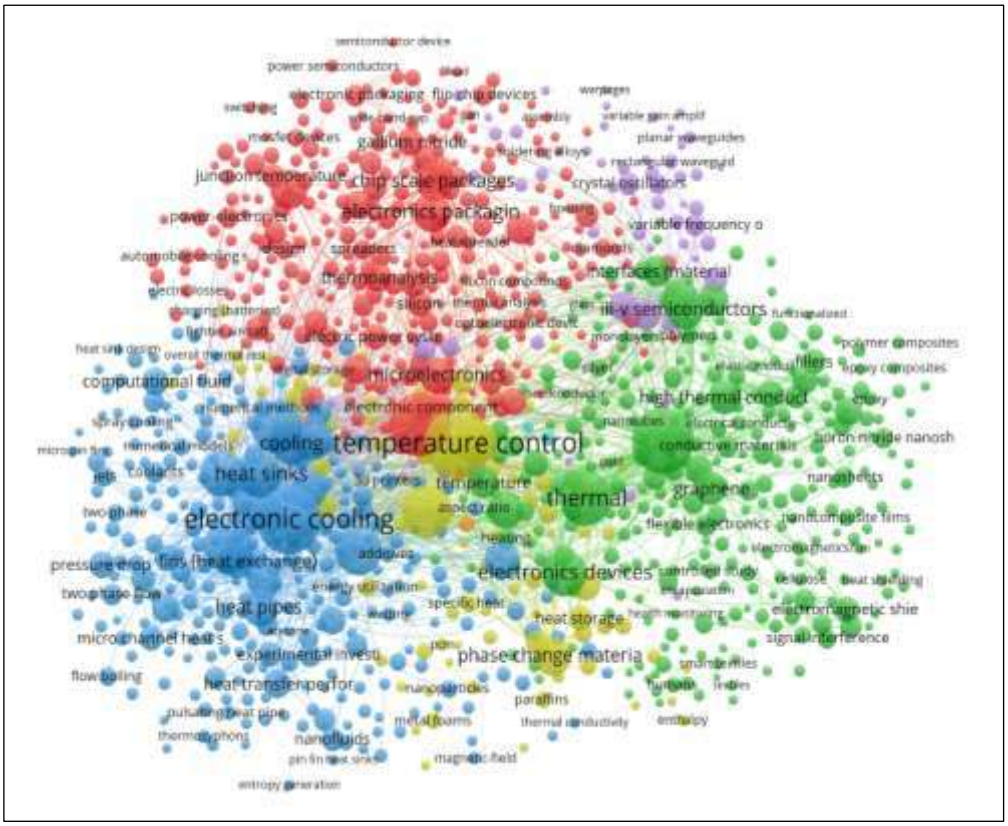


Figure 10: Diagram of network analysis using indexed keywords.
Source: Authors, (2026).

The network of different authors and their partnerships with co-authors is depicted in Figure 11. The author's minimum document count was limited to 5. Out of 13742 authors, it was possible to plot 261 of them on the graph. Data shows Yu Jinhong (21 articles, Citation 1730), Sun Rong (26 articles, Citation count 2006), Suresh Garimella (20 articles, Citation count 1224) and Avaram Bar Cohen (25 articles, Citation count 498) did a lot of collaborative work.

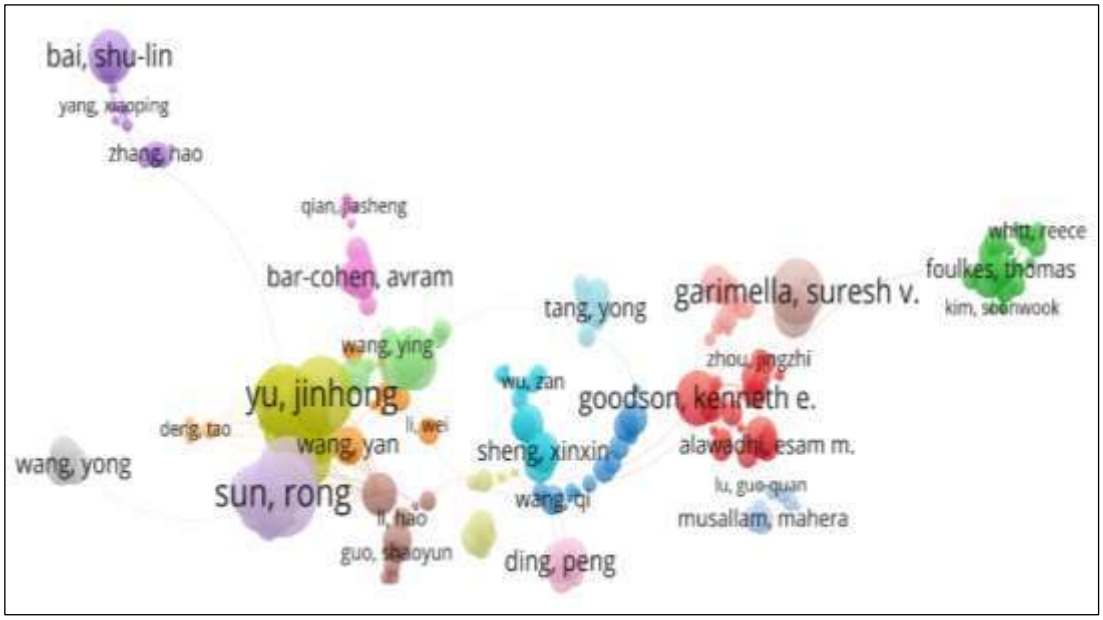


Figure 11: Diagram of the network analysis based on the authors' collaboration for TMED between 2000 and 2025.
Source: Authors, (2026).

Figure 12 shows that citations are continuously increasing over the years, which depicts the inclination of researchers in the domain. The year 2024 received the highest citations of 16563 for the total 616 documents published in Scopus database. Till June 2025, the citation count and published documents have reached 70% of previous year.

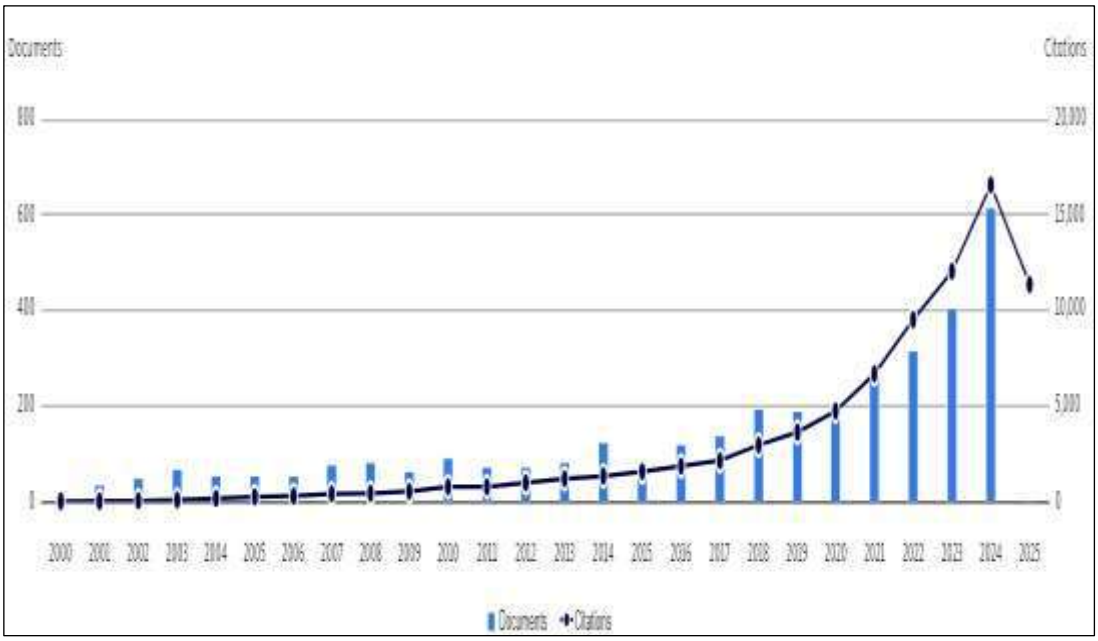


Figure 12: Diagram of the network analysis based on the citation trend on TMED between 2000 and 2025.
Source: Authors, (2026).

II.1.4 Studies on Thermal Management of Electronic Devices With Passive Cooling

(a) Research trends in the top ten countries

Keyword search using the words “Thermal Management of Electronic devices with passive cooling” returned contributions by 40 different countries in between 2000 to 2025. The search results were plotted for the to 10 top countries. The country-wise distribution of documents eliciting the number of research articles on TMED with passive cooling is shown in Figure 13, displays data on the 10 most productive countries.

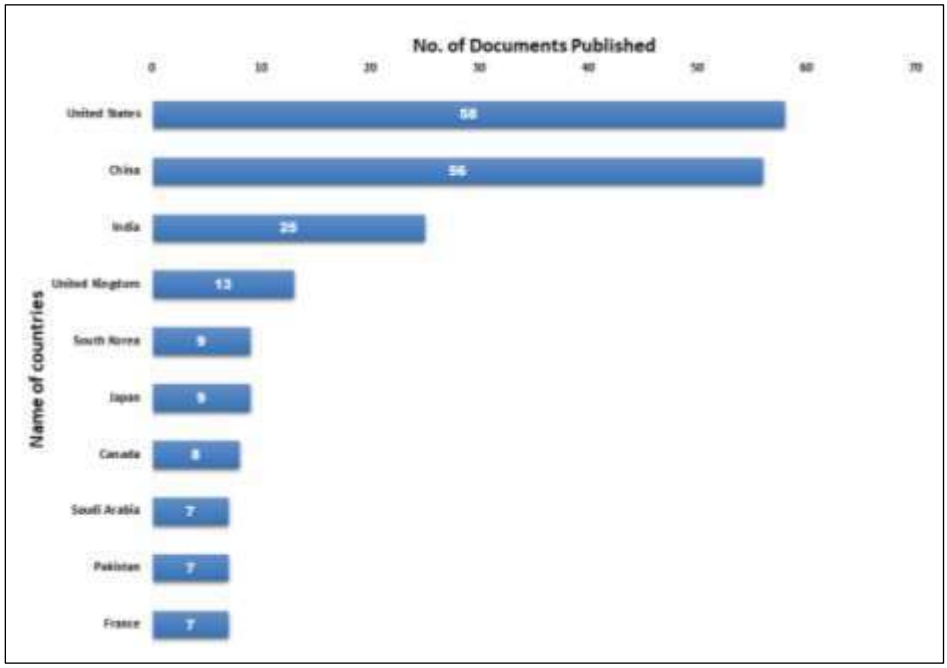


Figure 13: Country-wise research trend from 2000 to 2025.
Source: Authors, (2026).

China and the United States are almost equal in terms of publications, followed by India. The other Asian countries are behind in the domain publication.

(b) Collaboration between Countries

Data obtained from Scopus database using VOS viewer version 1.6.18. Bibliographic coupling analysis was performed to create network maps, as shown in Figure 14, connecting countries based on their collaborative publication. Authors from China, the United States and Saudi Arabia are collaborating more compared to other countries, as visible from the "link strength". It could be the reason why research is on an increasing trend in those countries.

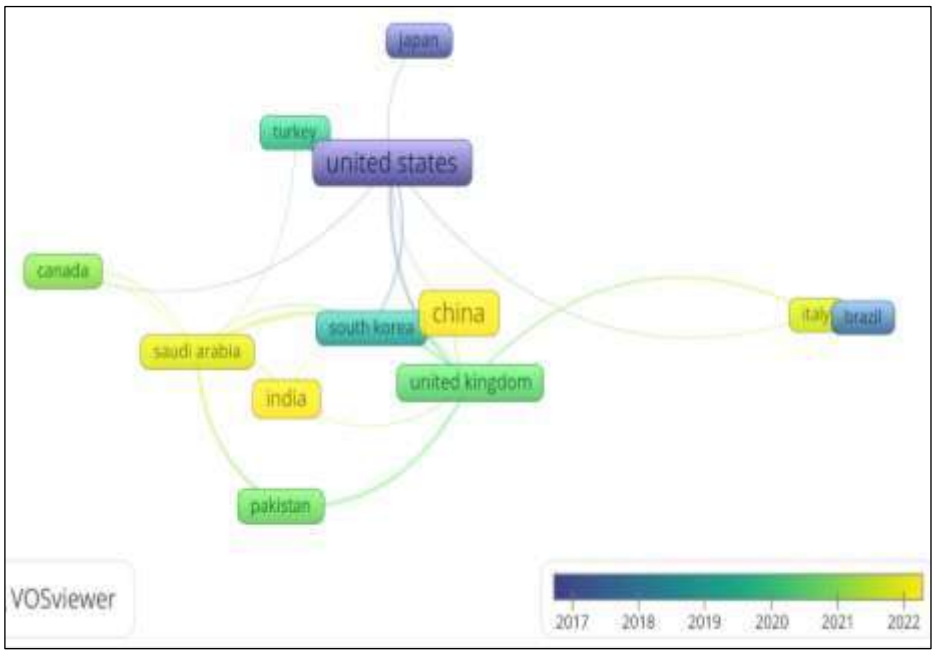


Figure 14: Network analysis diagram based on country keywords.
Source: Authors, (2026).

(c) Languages and Document Types

This section discusses the languages and document types seen in the subject matter. The documents were mainly published in four languages, namely English (219), Chinese (4). In terms of document type, book chapters (8), conference reviews (05), notes (1), articles (124), conference papers (76), reviews (10), and other document types contribute far less to this field of study as shown in Figure 15. The majority of the additional analysis was conducted using document-type articles, which are commonly found in these journals.

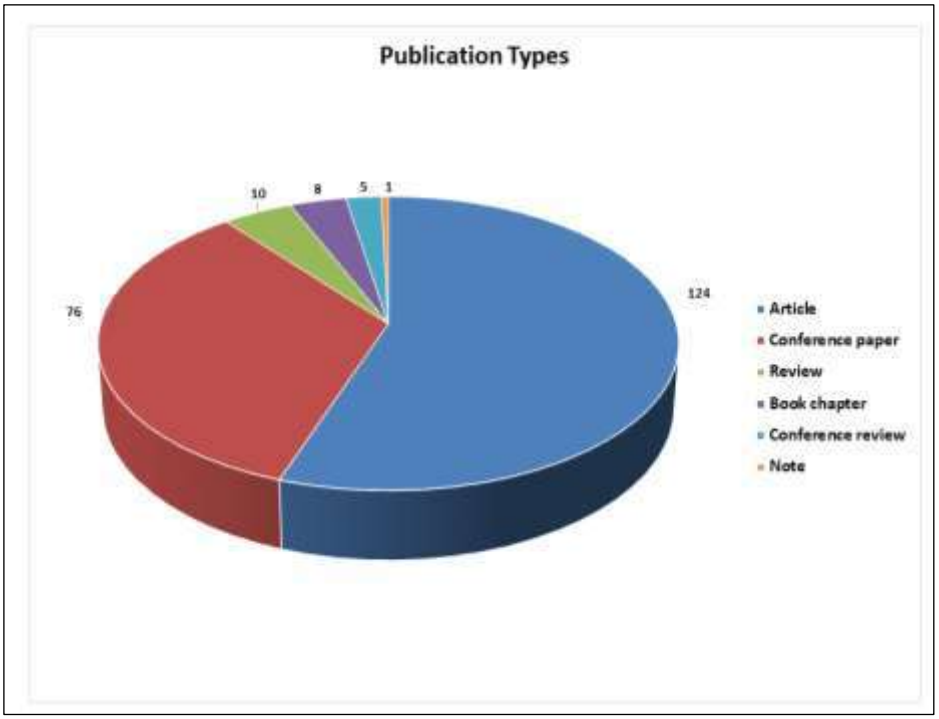


Figure 15: Document type distribution on TMED with passive cooling.
Source: Authors, (2026).

(d) Publication Trend

Figure 16 shows similar trend as that of earlier publications trend, signifying the interest of researchers in using passive cooling technologies like phase change materials, heat pipes in mitigating the challenges posed by electronic cooling.

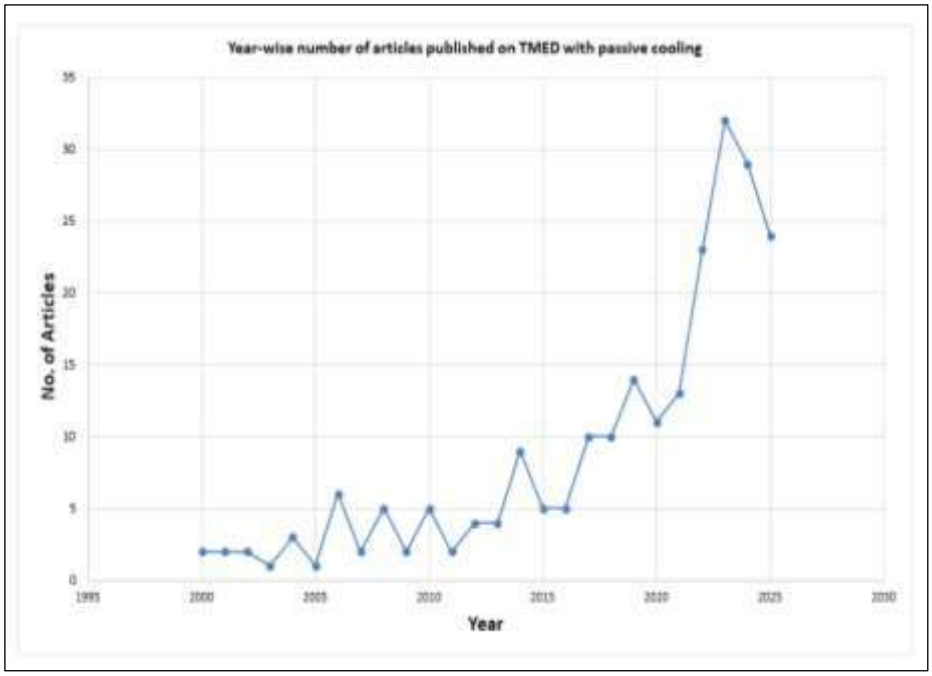


Figure 16: Publication trend year-wise on TMED with passive cooling.
Source: Authors, (2026).

(e) Subject area analysis

Figure 17 shows that significant research articles are published in other subject areas apart from the engineering domain, which shows that the cooling of electronic devices using passive cooling techniques holds many applications.

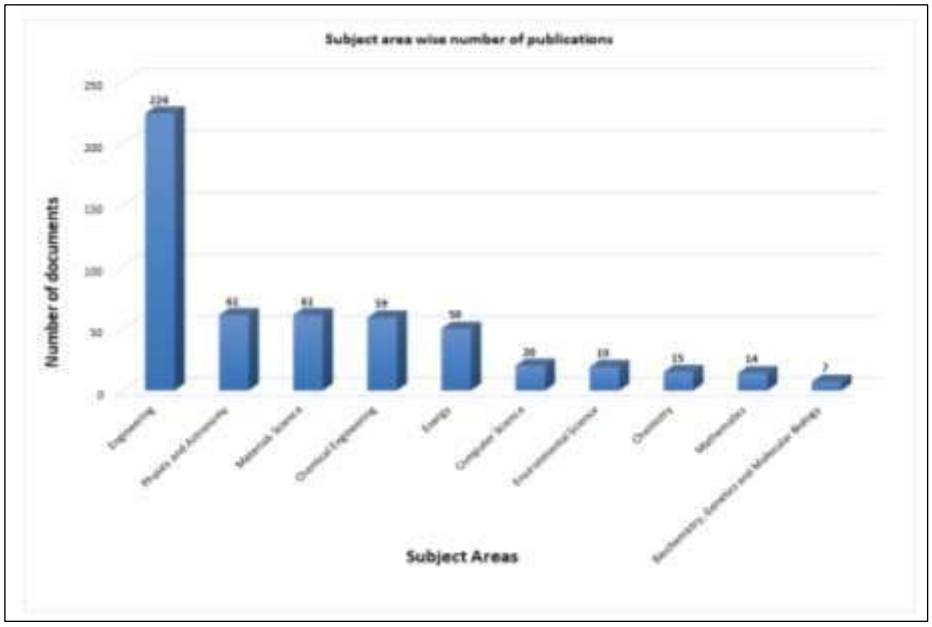


Figure 17: Subject area-wise on TMED with passive cooling.
Source: Authors, (2026).

(f) Keyword-Based Network Analysis and Author Collaboration

Figure 18 shows the network analysis performed on VOS viewer for cited author collaboration. The minimum number of documents of an author are considered as 2 and 66 authors meet the criteria out of 869. Figure 15 depicts the density visualization of 9 major clusters.

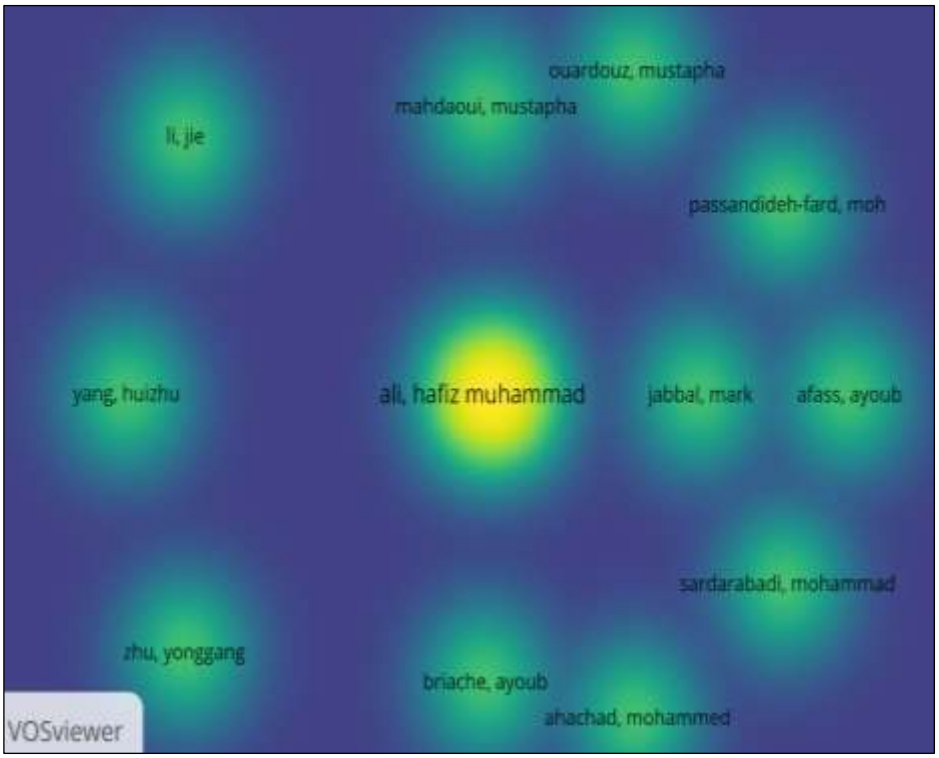


Figure 18: A network analysis diagram for TMED with passive cooling and vapour chamber from 2007 to 2024.
Source: Authors, (2026).

II 1.5 Studies On Thermal Management Of Electronic Devices With Passive Cooling And Vapour Chamber

(a) Research trends in the top ten countries

Referring to figure 19 fetched from Scopus data between 2010 to 2025, the number of documents published in this niche area is significantly lower. The United States is leading the way, followed by China and India. This shows there is a lot of scope for research in this domain.

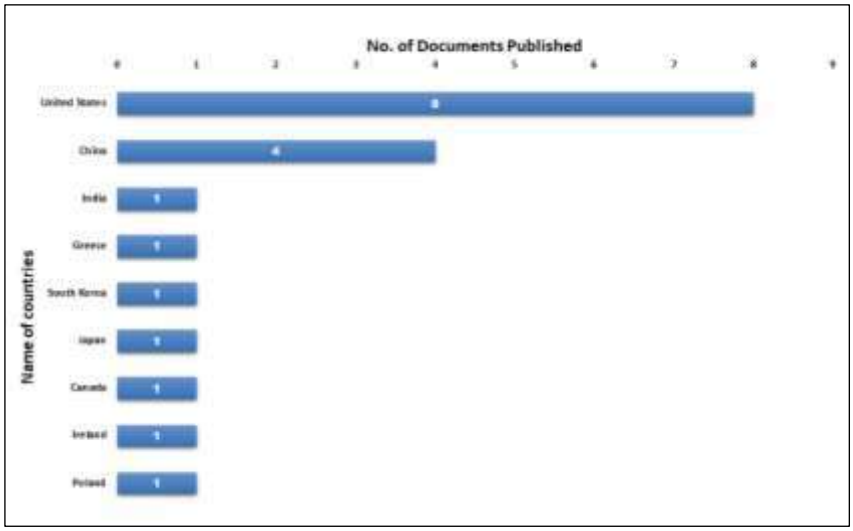


Figure 19: Country-wise research trend from 2010 to 2025.
Source: Authors, (2026).

(b) Collaboration between nations

For the term "Thermal management of electronic devices with passive cooling and vapour chamber " Figure 20 shows the network of different nations and their partnerships with other writers. A requirement of 1 document was set as a qualifying criterion in order to be reflected on the graph. United nation authors show more collaboration with other countries in publishing research articles in this domain.

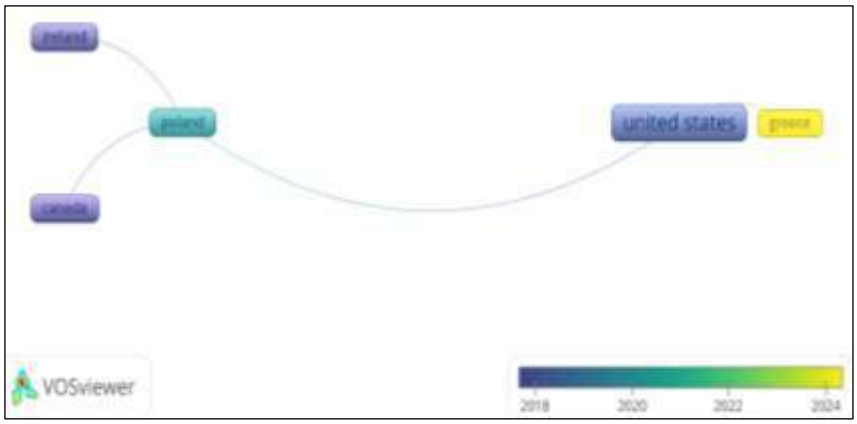


Figure 20: A country network analysis diagram for TMED with passive cooling and vapour chamber 2010 to 2024.
Source: Authors, (2026).

(c) Types of Documents

Figure 21 shows the research articles and the conference papers contribution is more as compare to other document types.

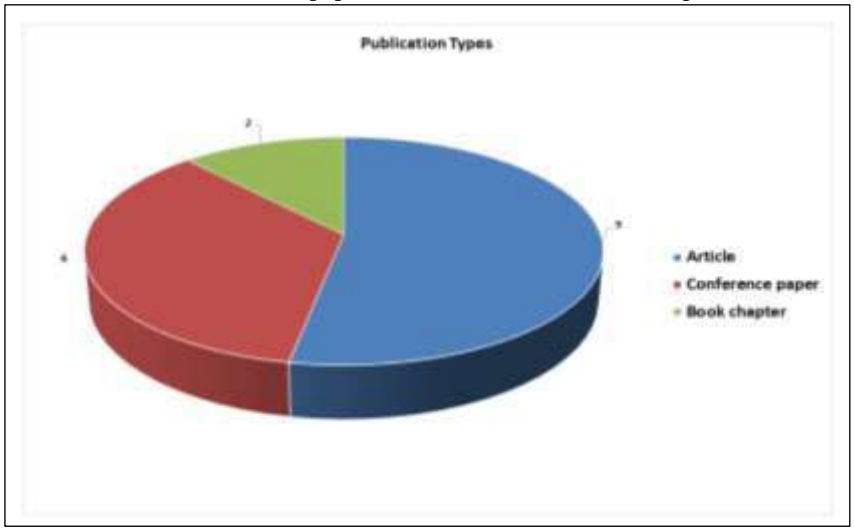


Figure 21: Document type distribution on TMED with passive cooling and vapour chambre.
Source: Authors, (2026).

(d) Publication Trend

Since only Scopus database is referred here, figure 22 shows the number of publication year wise is on the lower side. The count till 2016 is very less but the trend is gradually increasing from last two years.

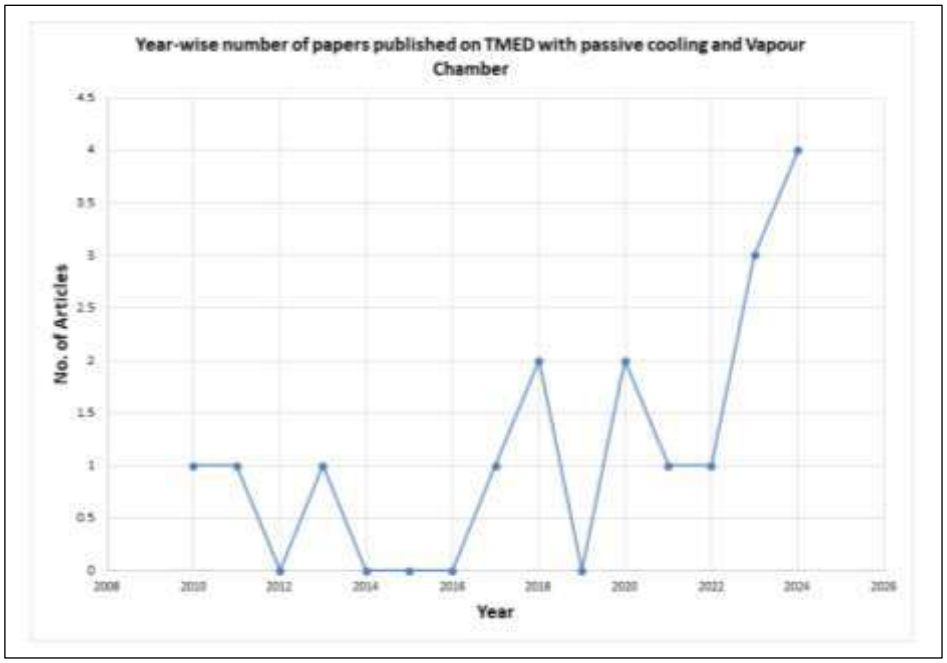


Figure 22: Publication trend year-wise on TMED with passive cooling and vapour chambre.
Source: Authors, (2026).

(e) Subject area analysis

Figure 23 shows the similar pattern as that of earlier two keywords. Physics and astronomy (7), Material science (6), Chemical engineering (3) follows the lead after engineering subject area due to relevant application in respective domain.

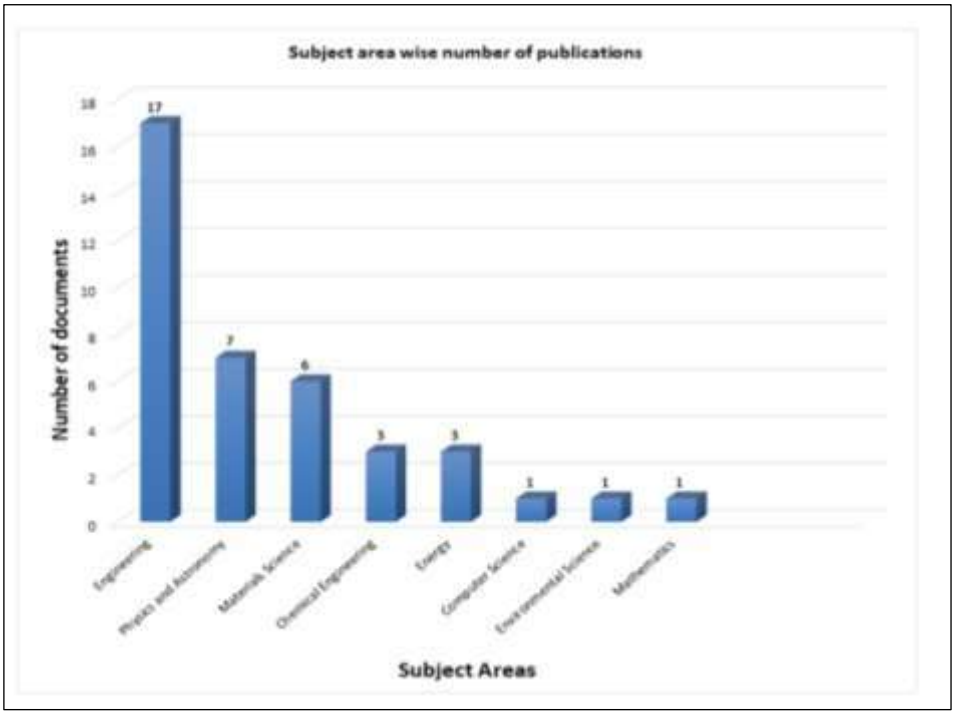


Figure 23: Subject area-wise on TMED with passive cooling and vapour chambre.
Source: Authors, (2026).

III COMPARATIVE ANALYSIS OF BIBLIOMETRIC TRENDS FOR ALL THE CATEGORIES

In this study, it is decided to do bibliometric analysis by three keyword search which narrow downs the research domains in following three categories:

Category -I: ----Thermal Management of Electronic devices
Category -II: --- Thermal Management of Electronic devices with passive cooling
Category -III: ---Thermal Management of Electronic devices with passive cooling and vapour chamber

The following section investigates all the three categories to arrive at meaningful conclusions

III.1 COUNTRY-WISE CONTRIBUTION

Overall, the number of publications started significantly rising after 2015 in the area of TMED. China (1380 documents, 32%) is at the top of graph, followed by United states may be due to the funding provided by universities and the allied industries. This helped the creation of value research and publication in the said countries. The Scopus data shows that more than 40 times Chinese universities are giving acknowledgement for sponsoring the required funds. Apart from China, there is hardly any country in Asia that is contributing to strengthening the domain. India, being a developing economy, shall focus on this research domain by extending a helping hand to the MSME sector in the creation of state-of-the-art manufacturing capabilities like 3D printing and rapid prototyping. The USA spends 3.59% of GDP on R&D in 2022, followed by China at 2.56%, while India falls far behind at 0.65%, according to the most recent data available.

With yearly growth rates of 8.7%, China exhibits the greatest development trajectory, whereas the US and India display more moderate growth trends [37]. The industry and academia shall come together to foster the research in such emerging areas [38] to bridge the gap between industrial needs and academic ecosystem. The limited availability of domestic funding compels researchers to pursue and depend on international collaboration. National scientific performance relies not solely on domestic R&D investments but is also affected by foreign investments from other countries [39]. Lower-income countries contribute minimally to Sustainable Development Goal (SDG) research, indicating a global misalignment [40] . This eventually affects the contribution of academia research in national/international environmental policies [41].

III.2 TYPE OF DOCUMENT

The data indicates that the number of research articles published (2365) is higher across all three categories, with 56%, 55% and 56% contributions in the first (2232 nos), second (124 nos) and third (9 nos) categories, respectively. The analysis reveals significant scope in the third category to explore the vapour chamber as a passive cooling medium in the thermal management of electronic devices, due to its superior thermal performance in a compact space.

The citation number of a publication tells volume about the significance and its penetration in research community. It also paves the path to take the current research ahead. The research trend in figure 12 clearly shows the spike in number of documents published as well as citations received are all time high in the year 2024 and similar trend is observed till mid-2025. This also indicates that there is growing need and market for research in this domain. It is very likely that the growing need of end user and progression of semiconductor technology is inviting more researchers to address the issues related to heat dissipation in compact spaces.

Data shows that engineering is not the only field wherein thermal issues related to electronic devices are pertinent. Physics and astronomy, owing to the development of semiconductor technology, and chemical engineering, owing to thermal management of control devices used in process industries and material science, owing to the development of lightweight and high thermal conductivity material development are following the similar trend of engineering as subject area. Overall, multidisciplinary research can lead to groundbreaking research in the area of TMED [41].

IV. CONCLUSIONS

The bibliometric analysis conducted based on Scopus publication data reveals that the thermal management of electronic devices is picking up pace in the market due to the digitalisation and use of compact high-power electronic equipment. The present study shows that China, followed by the United States of America, is leading the trend in terms of publication and research. This trend is because the funding agencies and the government policies are more oriented towards research and development. It is also observed that the author collaboration is more prevalent in these countries. This trend will attract more talent pool across the globe to seek higher education opportunities in these countries. Indian authors are gradually making progress in the area of thermal management and electronic cooling. In document types, the research articles are at the top, having more than 50% contribution for all three keywords chosen.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Nilesh Gaikwad, Narendra Deore and Mahadev Madgule.
Methodology: Nilesh Gaikwad, Narendra Deore and Mahadev Madgule.
Investigation: Nilesh Gaikwad, Narendra Deore and Mahadev Madgule.
Writing – Original Draft: Nilesh Gaikwad, Narendra Deore and Mahadev Madgule.
Writing – Review and Editing: Nilesh Gaikwad, Narendra Deore and Mahadev Madgule.
Resources: Nilesh Gaikwad, Narendra Deore and Mahadev Madgule.
Supervision: Nilesh Gaikwad, Narendra Deore and Mahadev Madgule.
Approval of the final text: Nilesh Gaikwad, Narendra Deore and Mahadev Madgule.

VII. REFERENCES

[1] N. Alam and M. Alam, “The Trend of Different Parameters for Designing Integrated Circuits from 1973 to 2019 and Linked to Moore’s Law,” Aust. J. Eng. Innov. Technol., vol. 2, no. 2, pp. 16–23, 2020, doi: 10.34104/ajeit.020.016023.

[2] G. Moore and O. N. M. S. Law, “Moore ’ s Law is live Moore ’ s,” 2005.

[3] F. Zhu, P. Xu, and J. Zong, “Moore’s Law: The potential, limits, and breakthroughs,” Appl. Comput. Eng., vol. 10, no. 1, pp. 307–315, 2023, doi: 10.54254/2755-2721/10/20230038.

[4] Z. He, Y. Yan, and Z. Zhang, “Thermal management and temperature uniformity enhancement of electronic devices by micro heat sinks: A review,” Energy, vol. 216, p. 119223, Feb. 2021, doi: 10.1016/J.ENERGY.2020.119223.

[5] A. R. Dhumal, A. P. Kulkarni, and N. H. Ambhore, “A comprehensive review on thermal management of electronic devices,” J. Eng. Appl. Sci., vol. 70, no. 1, pp. 1–18, 2023, doi: 10.1186/s44147-023-00309-2.

[6] H. Chen, T. Zhang, Q. Gao, J. Lv, H. Chen, and H. Huang, “Thermal management enhancement of electronic chips based on novel technologies,” Energy, vol. 316, p. 134575, Feb. 2025, doi: 10.1016/J.ENERGY.2025.134575.

[7] H. Lv, D. Jing, and W. Zhan, “Research on thermal management technology of system-level electronic integrated package,” J. Phys. Conf. Ser., vol. 2033, no. 1, 2021, doi: 10.1088/1742-6596/2033/1/012138.

[8] M. Azarifar, M. Arik, and J. Y. Chang, “Liquid cooling of data centers: A necessity facing challenges,” Appl. Therm. Eng., vol. 247, no. December 2023, p. 123112, 2024, doi: 10.1016/j.applthermaleng.2024.123112.

[9] H. Liang, W. Zhang, J. Huang, S. Yang, and P. Gupta, “Leveraging Hotspots and Improving Chip Reliability via Carbon Nanotube Grid Thermal Structure,” IEEE Trans. Very Large Scale Integr. Syst., vol. 23, no. 4, pp. 731–742, Apr. 2015, doi: 10.1109/TVLSI.2014.2321182.

[10] J. Shalf, “The future of computing beyond Moore ’ s Law Subject Areas :,” Philos. Trans. R. Soc., vol. 378, no. 20190061, pp. 1–14, 2020.

[11] Tanushree P, “Effect of using PCM as Cooling Medium in Different Types of Heat Sinks,” Int. J. Eng. Res., vol. V9, no. 08, pp. 201–206, 2020, doi: 10.17577/ijertv9is080103.

[12] H. Wang et al., “Bioinspired thermally conducting packaging for heat management of high performance electronic chips,” Commun. Eng., vol. 4, no. 1, pp. 2–10, 2025, doi: 10.1038/s44172-024-00338-6.

[13] H. Jouhara, D. Reay, R. McGlen, P. Kew, and J. McDonough, “Chapter 7 - Cooling of electronic components,” in Heat Pipes (Seventh Edition), Seventh Ed., H. Jouhara, D. Reay, R. McGlen, P. Kew, and J. McDonough, Eds., Butterworth-Heinemann, 2024, pp. 271–306. doi: https://doi.org/10.1016/B978-0-12-823464-8.00006-1.

[14] N. Sharifi, H. Shabgard, C. Millard, and U. Etufugh, “Hybrid Heat Pipe-PCM-Assisted Thermal Management for Lithium-Ion Batteries,” Batteries, vol. 11, no. 2, 2025, doi: 10.3390/batteries11020064.

[15] S.-J. Kim, J.-S. Lee, and S.-H. Rhi, “Heat Pipe Embedded Battery Cooling System for Future Electric Vehicle,” Batteries, vol. 11, no. 4, 2025, doi: 10.3390/batteries11040164.

[16] M. A. Gómez, R. Bellas, A. González-Gil, A. Cacabelos, and A. Larrañaga, “Thermal study of a passive cooling device operating through a bubble lifting CLTPT of NOVEC 7000 with a two-fluid condenser,” Int. J. Heat Mass Transf., vol. 177, 2021, doi: 10.1016/j.ijheatmasstransfer.2021.121530.

[17] A. Ghanbarpour, M. J. Hosseini, A. A. Ranjbar, M. Rahimi, R. Bahrampoury, and M. Ghanbarpour, “Evaluation of heat sink performance using PCM and vapor chamber/heat pipe,” Renew. Energy, vol. 163, pp. 698–719, 2021, doi: 10.1016/j.renene.2020.08.154.

[18] A. A. Litvintceva, N. I. Volkov, N. I. Vorogushina, V. A. Moskovskikh, and V. V. Cheverda, “The effect of operating parameters on the heat transfer in the heat pipe,” J. Phys. Conf. Ser., vol. 2119, no. 1, 2021, doi: 10.1088/1742-6596/2119/1/012088.

[19] Y. Luo, Y. Tang, X. Zhang, H. Wang, G. Zhou, and P. Bai, “A novel composite vapor chamber for battery thermal management system,” Energy Convers. Manag., vol. 254, no. February, p. 115293, 2022, doi: 10.1016/j.enconman.2022.115293.

[20] J. Zhao, M. Du, Z. Zhang, and X. Ling, “Thermal management strategy for electronic chips based on combination of a flat-plate heat pipe and spray cooling,” Int. J. Heat Mass Transf., vol. 181, p. 121894, Dec. 2021, doi: 10.1016/J.IJHEATMASSTRANSFER.2021.121894.

[21] A. Shafieian, M. Khiadani, and A. Nosrati, “Strategies to improve the thermal performance of heat pipe solar collectors in solar systems: A review,” Energy Convers. Manag., vol. 183, pp. 307–331, Mar. 2019, doi: 10.1016/J.ENCONMAN.2018.12.115.

[22] M. Bulut, S. G. Kandlikar, and N. Sozbir, “A Review of Vapor Chambers,” Heat Transf. Eng., vol. 40, no. 19, pp. 1551–1573, 2019, doi: 10.1080/01457632.2018.1480868.

[23] D. Reay, R. J. McGlen, and P. A. Kew, “Heat Pipes: Theory, Design and Applications,” 2006.

[24] T. Zhao and Z. Hu, “Opportunities and Challenges for the Development of Ultra-Thin Heat Pipe,” Front. Energy Res., vol. 10, no. June, pp. 1–3, 2022, doi: 10.3389/fenrg.2022.923454.

[25] P. Nemec, A. Čaja, and M. Malcho, “Mathematical model for heat transfer limitations of heat pipe,” Math. Comput. Model., vol. 57, no. 1–2, pp. 126–136, 2013, doi: 10.1016/j.mcm.2011.06.047.

[26] M. M. Farid, A. M. Khudhair, S. A. K. Razack, and S. Al-Hallaj, “A review on phase change energy storage: Materials and applications,” Energy Convers. Manag., vol. 45, no. 9–10, pp. 1597–1615, 2004, doi: 10.1016/j.enconman.2003.09.015.

[27] A. Sharma, V. V Tyagi, C. R. Chen, and D. Buddhi, “Review on thermal energy storage with phase change materials and applications,” Renew. Sustain. Energy Rev., vol. 13, no. 2, pp. 318–345, 2009, doi: https://doi.org/10.1016/j.rser.2007.10.005.

[28] A. B. Etemoglu and M. Can, “Classification of geothermal resources in Turkey by exergy analysis,” Renew. Sustain. Energy Rev., vol. 11, no. 7, pp. 1596–1606, 2007, doi: 10.1016/j.rser.2006.01.001.

[29] A. Mohammed Adham, N. Mohd-Ghazali, and R. Ahmad, “Thermal and hydrodynamic analysis of microchannel heat sinks: A review,” Renew. Sustain. Energy Rev., vol. 21, pp. 614–622, 2013, doi: 10.1016/j.rser.2013.01.022.

[30] Z. Yuan, G. Vaartstra, P. Shukla, S. Reda, E. Wang, and A. K. Coskun, “Modeling and Optimization of Chip Cooling with Two-Phase Vapor Chambers,” Proc. Int. Symp. Low Power Electron. Des., vol. 2019-July, pp. 1–6, 2019, doi: 10.1109/ISLPED.2019.8824965.

[31] S. M. Kim and I. Mudawar, “Universal approach to predicting saturated flow boiling heat transfer in mini/micro-channels – Part II. Two-phase heat transfer coefficient,” Int. J. Heat Mass Transf., vol. 64, pp. 1239–1256, 2013, doi: 10.1016/j.ijheatmasstransfer.2013.04.014.

[32] R. Wen, W. Liu, X. Ma, and R. Yang, “Coupling Droplets/Bubbles with a Liquid Film for Enhancing Phase-Change Heat Transfer,” iScience, vol. 24, p. 102531, 2021, doi: 10.1016/j.isci.2021.102531.

[33] H. T. N. Ho and H. T. Luong, “Research trends in cybercrime victimization during 2010–2020: a bibliometric analysis,” SN Soc. Sci., vol. 2, no. 1, p. 4, 2022, doi: 10.1007/s43545-021-00305-4.

[34] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, “How to conduct a bibliometric analysis: An overview and guidelines,” J. Bus. Res., vol. 133, no. May, pp. 285–296, 2021, doi: 10.1016/j.jbusres.2021.04.070.

[35] G. Chen and L. Xiao, “Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods,” J. Informetr., vol. 10, no. 1, pp. 212–223, 2016, doi: 10.1016/j.joi.2016.01.006.

[36] U. A. Bukar, M. S. Sayeed, S. F. A. Razak, S. Yogarayan, O. A. Amodu, and R. A. R. Mahmood, “A method for analyzing text using VOSviewer,” MethodsX, vol. 11, no. May, p. 102339, 2023, doi: 10.1016/j.mex.2023.102339.

[37] A. Pratap, “A look at R&D spending plans across the globe.,” C&EN Global Enterprise, p. 103, 2025. doi: https://doi.org/10.1021/cen-10301-feature5.

[38] F. Ahmed, M. T. Fattani, S. R. Ali, and R. N. Enam, “Strengthening the Bridge Between Academic and the Industry Through the Academia-Industry Collaboration Plan Design Model,” Front. Psychol., vol. 13, no. July 2016, pp. 1–11, 2022, doi: 10.3389/fpsyg.2022.875940.

[39] L. Miao, V. Larivière, F. Wang, Y.-Y. Ahn, and C. R. Sugimoto, “Cooperation and interdependence in global science funding,” 2023, [Online]. Available: http://arxiv.org/abs/2308.08630

[40] H. Confraria, T. Ciarli, and E. Noyons, “Countries’ research priorities in relation to the Sustainable Development Goals,” Res. Policy, vol. 53, no. 3, p. 104950, 2024, doi: 10.1016/j.respol.2023.104950.

[41] A. Singh, S. H. Qasim, and S. S. Singh, “Importance of research in the development of the nation,” Int. J. All Res. Educ. Sci. Methods, vol. 9, no. 7 July, pp. 5–9, 2021.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 763-773. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3057

RESEARCH ARTICLE OPEN ACCESS

EXPERIMENTAL STUDY ON NR–SBR RUBBER COMPOSITE WITH CARBON BLACK FILLER FOR INDUSTRIAL TRACK PAD APPLICATIONS

Farit Hendro Wibowo*[1], Sujito Sujito [2] and Muhammad Rodhi Faiz[3]

[1, 2, 3]Electronics and Informatics Engineering University of Malang, Malang, Indonesia.

[1]https://orcid.org/0000-0003-3567-3822, [2]http://orcid.org/0000-0001-5917-306X, [3]https://orcid.org/0000-0002-6684-739X

Email: *farit.hendro.2305349@students.um.ac.id, sujito.ft@um.ac.id, mohamad.rodhi.ft@um.ac.id

ARTICLE INFO

Article History

Received: December 5, 2025

Revised: January 10, 2026

Accepted: January 15, 2026

Published: February 28, 2026

Keywords:

Natural Rubber (NR),
Styrene-Butadiene Rubber (SBR),
Carbon Black,
Rubber Composite,
Mechanical Properties,
Industrial Track Pad.

ABSTRACT

This study investigates the performance of NR–SBR (80/20) rubber composites reinforced with Carbon Black N330 for use in industrial track pad applications. Five formulations containing 65, 70, 75, 80, and 85 phr of carbon black were prepared and evaluated. Mechanical properties such as hardness, tensile strength, elongation at break, abrasion resistance, and dynamic behavior were measured according to relevant ASTM standards. In addition, field testing was conducted using a tracked vehicle over a 15 km route to assess real-world wear behavior. The results show that carbon black content significantly influences stiffness, strength, and abrasion resistance. Among all formulations, the 75 phr compound demonstrated the best balance of hardness, tensile strength, and wear performance, making it a suitable candidate for industrial track pad applications. These findings provide useful insights for developing locally produced rubber composites with improved durability and cost efficiency.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Rubber materials and their composites have long played a vital role in industrial and transportation applications due to their excellent flexibility, damping ability, and wear resistance. Natural rubber (NR), as one of the most versatile elastomers, provides superior tensile strength, resilience, and fatigue resistance, which make it highly suitable for dynamic load conditions[1], [2]. However, NR alone exhibits limitations in thermal aging, ozone resistance, and oil compatibility, which can reduce its long-term durability in demanding industrial environments. Styrene-butadiene rubber (SBR), on the other hand, offers improved aging stability, abrasion resistance, and processability, making it an ideal synthetic counterpart to be blended with NR. The combination of these two elastomers, when properly compounded, can produce a hybrid rubber material that integrates the elasticity of NR with the toughness and wear resistance of SBR.

To achieve optimal mechanical behavior, reinforcing fillers such as carbon black are commonly incorporated into the blend to enhance stiffness, tensile strength, and resistance to mechanical degradation[3], [4]. The role of carbon black as a reinforcing filler in rubber compounds is critical for achieving the required performance in industrial applications. Carbon black enhances stress transfer efficiency between the polymer chains by creating strong filler–matrix interactions and improving cross-linking density during vulcanization. The dispersion level and concentration of carbon black determine the compound's hardness, elongation, and abrasion properties, which must be precisely balanced to meet specific functional demands. Previous studies have shown that increasing the filler concentration generally improves stiffness and wear resistance, but excessive filler content can lead to filler agglomeration and brittleness[5], [6].

Therefore, identifying the optimum filler loading is essential for maintaining both flexibility and durability. For applications such as industrial track pads and high-load contact components, the mechanical balance between hardness and elasticity is vital to ensure long service life under cyclic deformation and impact conditions. In industrial track systems, such as heavy vehicles or tracked machinery, the pad shoe or track pad serves as a primary interface between the moving track and the ground surface. These components experience repetitive compressive, shear, and abrasive forces during operation, requiring materials that combine high hardness with energy-absorbing capacity.

Conventional materials used for these parts are often imported and expensive, posing challenges for maintenance and supply chain sustainability[7-9]. Developing an NR–SBR rubber composite using locally available raw materials can therefore provide significant economic and strategic benefits. The optimization of mechanical performance through carbon black modification not only improves product quality but also supports the development of domestic material technologies capable of replacing imported rubber components. This aligns with the broader industrial objective of self-reliance and sustainable manufacturing[3], [10], [11]. Numerous research efforts have explored NR–SBR blends with different fillers, focusing on mechanical properties such as tensile strength, abrasion resistance, and dynamic behavior.

Studies by Hasan (2013) and Kinasih (2016) demonstrated that the inclusion of carbon black substantially improves hardness and tensile performance in NR–SBR compounds. However, limited studies have addressed the actual application of these materials in heavy-duty or dynamic systems such as track pads used in armored or industrial vehicles. Most investigations remain laboratory-based, with little validation through real-world load testing. This gap indicates the need for further experimental studies that not only examine the physical and mechanical behavior of NR–SBR composites under controlled conditions but also verify their field performance to ensure reliability under industrial stress environments[12], [13]. Based on these considerations, the present study focuses on experimentally investigating the influence of carbon black filler variation on the mechanical and physical properties of NR–SBR rubber composites intended for industrial track pad applications.

The research aims to determine the optimal filler loading that achieves the best balance between hardness, tensile strength, elongation, and abrasion resistance. Laboratory tests following ASTM and DIN standards were complemented by field dynamic load validation to simulate real working conditions. The integration of both laboratory and field data provides a comprehensive understanding of material performance, offering valuable insights into the development of durable, high-strength rubber materials for industrial and defense-related machinery. Ultimately, this research contributes to advancing local rubber technology innovation and supports the long-term goal of industrial independence and reliability in critical mechanical components.

II. THEORETICAL REFERENCE

Rubber materials are highly elastic polymers whose mechanical performance depends on molecular structure and cross-link density. Natural rubber (NR) is composed mainly of cis-1,4-polyisoprene chains that exhibit outstanding elasticity and tensile strength, which make it suitable for dynamic load-bearing components. The repeating unit of cis-1,4-polyisoprene, illustrated schematically in Fig. 1, shows alternating single and double bonds along the carbon backbone with pendant –CH_3 and –H groups that enable flexible molecular motion. However, NR is susceptible to oxidation and thermal degradation at elevated temperatures, limiting its long-term durability. To overcome these drawbacks, NR is often blended with styrene-butadiene rubber (SBR), whose copolymer chain contains styrene and butadiene units, providing improved abrasion and ageing resistance[14]. A conceptual structure of SBR is shown in Fig. 2, where alternating styrene-aromatic rings and butadiene segments contribute to hardness and wear resistance while maintaining adequate flexibility.

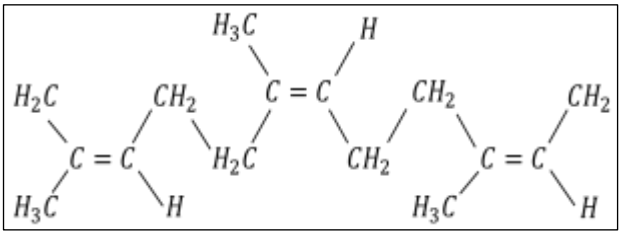


Figure 1: Schematic structure of cis-1,4-polyisoprene (NR).
Source: Authors, (2026).

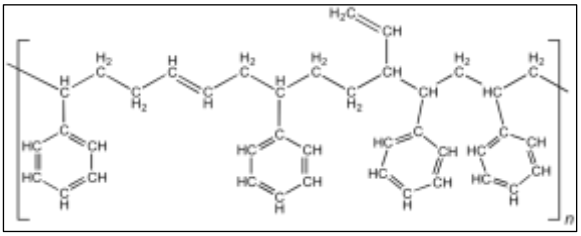


Figure 2: Schematic structure of styrene–butadiene copolymer (SBR).
Source: Authors, (2026).

The reinforcement of rubber compounds commonly involves the incorporation of fillers such as carbon black to enhance stiffness, tensile strength, and abrasion resistance. Carbon black particles interact physically and chemically with the polymer matrix, forming a semi-continuous filler network that restricts molecular mobility and increases modulus. As shown schematically in Fig. 3, these filler–matrix interactions improve stress transfer during deformation and contribute to energy dissipation under cyclic loading[15]. Nevertheless, excessive filler loading can cause agglomeration and stress concentration, which reduce elongation and resilience. Therefore, determining the optimum carbon-black concentration is essential to balance hardness, flexibility, and energy absorption—particularly for applications requiring dynamic impact resistance, such as industrial track pads.

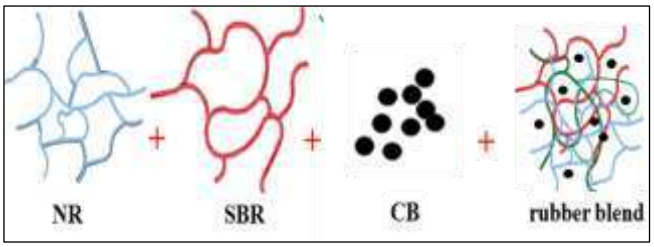


Figure 3: Filler–matrix interaction in NR–SBR composite.
Source: Authors, (2026).

The vulcanization process also has a significant influence on the final properties of NR–SBR composites. During vulcanization, sulfur cross-links form between polymer chains, converting the raw elastomer into a resilient and dimensionally stable material. The degree of cross-linking directly affects hardness, tensile strength, and resistance to permanent deformation. A schematic representation of the sulfur cross-linking reaction is illustrated in Fig. 4, showing how inter-chain bridges stabilize the polymer network. Studies by Chollakup et al. demonstrated that well-dispersed carbon black combined with optimal curing conditions can simultaneously improve tensile strength and abrasion resistance[16]. The mechanical balance between NR and SBR ratios is thus crucial: higher NR content enhances elasticity and fatigue life, while higher SBR content improves wear performance under repeated contact stress.

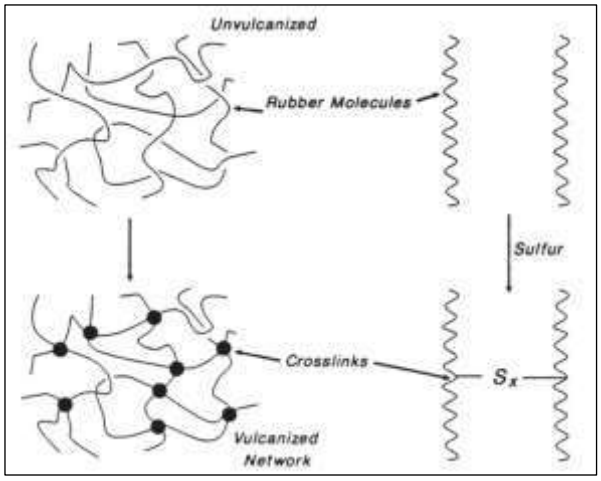


Figure 4: Sulfur cross-linking mechanism during vulcanization.
Source: Authors, (2026).

The viscoelastic behavior of filled rubber can be evaluated using dynamic mechanical analysis (DMA). The fundamental relation between storage modulus (E′) and loss modulus (E″) is expressed in Eq. (1), where the ratio tan δ = E″/E′ represents the damping capability of the material[17].

$$\tan\delta = \frac{E''}{E'} \tag{1}$$

A high storage modulus indicates stiffness, whereas a higher loss factor implies better vibration damping. The schematic in Fig. 5 summarizes this viscoelastic response for a filled elastomer under cyclic strain. For industrial track-pad applications, an optimal combination of hardness, elongation, and damping is required to achieve traction, comfort, and thermal stability. Hybrid fillers such as silica or graphene have been explored, but carbon black remains the most cost-effective and reliable reinforcement for heavy-duty elastomer systems[18]. Finally, theoretical predictions must be validated through experimental and field evaluations. In the present research, NR–SBR composites with varying carbon-black fillers were formulated and tested to determine their mechanical (hardness, tensile strength, elongation) and physical (abrasion resistance) properties.

The theoretical concepts described above—polymer microstructure (Fig. 1 – 2), filler–matrix reinforcement (Fig. 3), and viscoelastic behavior (Eq. 1, Fig. 5)—serve as the foundation for interpreting the experimental findings. Field testing under dynamic loads, as shown schematically in Fig. 6, provides practical confirmation of laboratory results and verifies the suitability of the developed composite for industrial track-pad applications. Thus, the theoretical framework integrates molecular structure, filler mechanics, and dynamic behavior to support the experimental study presented in this paper.

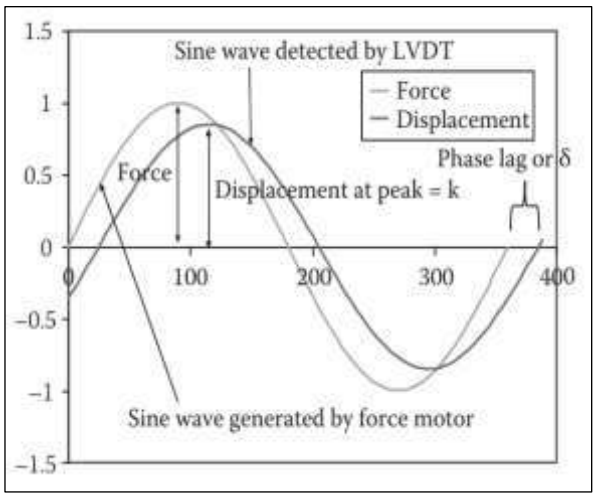


Figure 5: Typical Viscoelastic Response of Filled Rubber Under Cyclic Strain.
Source: [17].

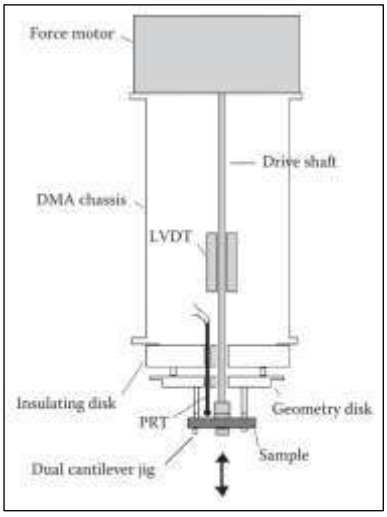


Figure 6: Simplified Schematic Of Dynamic Loading Conditions.
Source: [17].

III. MATERIALS AND METHODS

III.1 MATERIALS

The materials used in this study consisted of Natural Rubber (NR, SIR-20 grade) and Styrene–Butadiene Rubber (SBR 1502), which served as the primary elastomer blend. Carbon black N330 was selected as the reinforcing filler due to its well-established ability to enhance strength, abrasion resistance, and durability in rubber composites. Additional compounding agents included zinc oxide (ZnO) and stearic acid as activators, sulfur as the vulcanizing agent, processing oil as a softener, and CBS (N-cyclohexyl-2-benzothiazolesulfenamide) as the accelerator[19]. All materials were used in raw form without further purification. The molecular structures of NR and SBR are illustrated in Figures 1 and 2 to highlight their chain configurations relevant to elastomeric performance.

III.2 COMPOUND FORMULATION

Five compound formulations, labeled A through E, were prepared with varying carbon black loadings to evaluate the effect of filler concentration on mechanical strength and abrasion behavior. All ingredients were measured in parts per hundred rubber (phr), and the compositions of each formulation are summarized in Table 1. These formulations were designed to represent incremental increases in filler reinforcement, allowing systematic observation of changes in compound performance as the carbon black content increased.

Table 1: Formulation of NR–SBR Rubber Compounds (phr).

Material	Model A (phr)	Model B (phr)	Model C (phr)	Model D (phr)	Model E (phr)
Natural Rubber (NR)	80	80	80	80	80
SBR	20	20	20	20	20
Carbon Black (N330)	65	70	75	80	85
ZnO	5	5	5	5	5
Peptizer	2	2	2	2	2
Stearic Acid (STA)	1	1	1	1	1
6PPD	2	2	2	2	2
TMQ	1	1	1	1	1
Sulfur	3	3	3	3	3
TMTD	0.3	0.3	0.3	0.3	0.3
Accelerator (CBS)	1.2	1.2	1.2	1.2	1.2
Total	180.5	185.5	190.5	195.5	200.5

Source: Authors, (2026).

III.3 PREPARATION OF MATERIALS FOR RUBBER COMPOUNDING.

All raw materials were prepared according to the NR–SBR compound formulation presented in Table 1. The base elastomers—natural rubber (NR) and styrene–butadiene rubber (SBR)—were conditioned at room temperature to stabilize their viscoelastic properties before mixing. Carbon black (N330), used as the primary reinforcing filler, was weighed precisely along with processing oil, ZnO, stearic acid, TMQ, 6PPD, sulfur, and the accelerator system to ensure accurate phr ratios. The overall preparation workflow is illustrated in Figure 3, which outlines the sequential steps from material weighing to compound sheet formation.

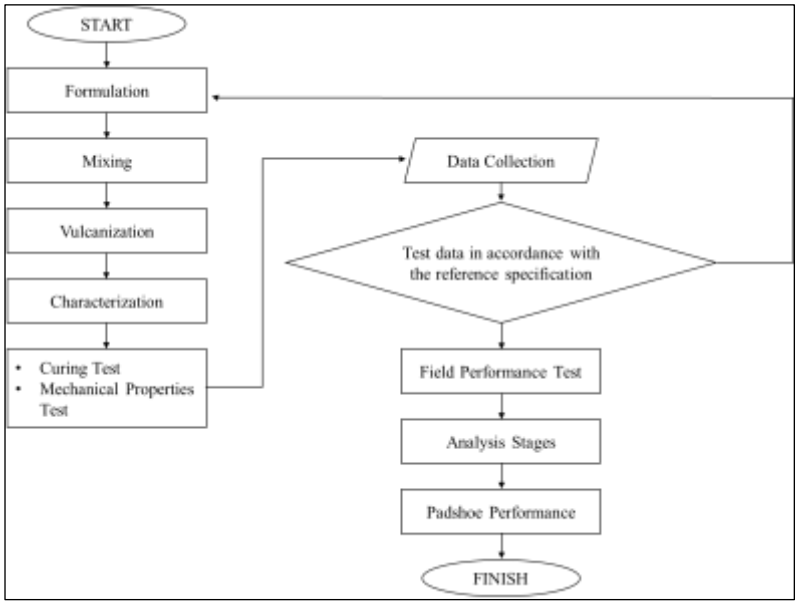


Figure 7: Flowchart of research methods.
Source: Authors, (2026).

The compounding process was carried out using a laboratory two-roll mill. The NR–SBR blend was first masticated until uniform plasticity was achieved. Carbon black was incorporated gradually to minimize agglomeration and promote homogeneous filler dispersion within the polymer matrix. ZnO and stearic acid were added as activators, followed by antioxidants (TMQ and 6PPD) and processing oil to enhance processability. The curing agents—sulfur and accelerators—were introduced during the final mixing stage under controlled roll temperature to prevent premature vulcanization[20]. After mixing, the compound sheets were removed, cooled, and rested for 24 hours to allow internal stress relaxation before the vulcanization stage[20]. These prepared NR–SBR compounds served as the base material for fabricating industrial track pad specimens. The complete process pathway—starting from raw material preparation, rubber compounding, sheet formation, and track pad molding—is presented schematically in Figure 3.

III.4 VULCANIZATION PROCESS.

The vulcanization process began after completing the compounding stage, in which NR and SBR were masticated on a two-roll mill and sequentially mixed with activators, carbon black filler, process oil, and the sulfur–accelerator curing system until a uniform rubber compound was obtained. The mixed compound was sheeted and allowed to rest to ensure stable filler dispersion before curing. Vulcanization was then carried out using a hydraulic hot press at 150 °C and 10 MPa. The optimum cure time (t90) for each formulation was determined using a rheometer in accordance with ASTM D2084 and applied during the press-curing stage to achieve the required cross-link density. The cured sheets were cooled under pressure and conditioned at room temperature for subsequent laboratory testing[21]. For track pad fabrication, the vulcanized compound was further molded into its final geometry using a steel compression mold. Pre-cut rubber preforms were placed into the mold cavity and subjected to the same optimized temperature and pressure conditions to ensure complete vulcanization throughout the track pad body. The molded track pads were cooled in the mold, demolded, trimmed, and prepared for dynamic field testing.

III.5 MECHANICAL TESTING.

Mechanical testing of the NR–SBR/carbon black composites was conducted following the workflow illustrated in Figure 8, covering rheometric analysis, controlled vulcanization, specimen conditioning, and a complete mechanical and dynamic performance evaluation. The optimum curing time (t_{90}) for each formulation was determined using a Moving Die Rheometer (MDR Professional, MonTech, Germany) at 140 °C in accordance with ASTM D5289. These t_{90} values were used as the basis for the hot-press vulcanization process, performed at 150 °C using a hydraulic press to ensure proper cross-link development. After curing, all samples were cooled under pressure and conditioned at room temperature for 24 hours.

Mechanical and physical characterization was carried out in accordance with the relevant ASTM standards. Tensile strength, tear strength, and elongation at break were measured using a Universal Testing Machine (Gotech AI-7000-S, 25 kN) following ASTM D412. Shore A hardness was assessed using a Mitutoyo durometer in accordance with ASTM D2240, with measurements taken at five distinct points per specimen. Density testing was performed using a laboratory balance (A&D GR-200) equipped with a density determination kit, while rebound resilience was evaluated using a Gibitre rebound resilience tester to quantify elastic recovery.

Abrasion resistance was examined using a DIN Abrasion Tester (AT150) according to ASTM D5963, producing volume-loss values representative of wear behavior under dry sliding conditions. Compression set testing was conducted following ASTM D395, where samples were compressed for 22 hours at 70 °C and subsequently stabilized for one hour at room temperature to assess permanent deformation characteristics[22]. The dispersion quality of carbon black within the NR–SBR matrix was evaluated using a Disperse Tester 3000 (MonTech) to ensure uniform filler distribution and mixing consistency. Dynamic mechanical properties were examined using a Dynamic Mechanical Analyzer (DMA7100, HITACHI High-Tech). Measurements were conducted in tensile mode across a frequency range of 0.1–10 Hz and a temperature sweep from −70 °C to 110 °C, with a heating rate of 2 °C/min.

Specimens measuring approximately 20 × 9 × 2 mm³ were used to evaluate the viscoelastic responses, including storage modulus (E′), loss modulus (E″), and damping characteristics (tan δ). All laboratory evaluations were complemented by field-simulation testing on fabricated track pad specimens to assess wear progression, crack formation, and surface degradation under dynamic loading. The complete characterization pathway—from rheometer-based curing evaluation to mechanical, physical, and dynamic testing—is summarized in Figure 8, representing the integrated assessment framework used to evaluate the NR–SBR/carbon black composites for industrial track pad applications.

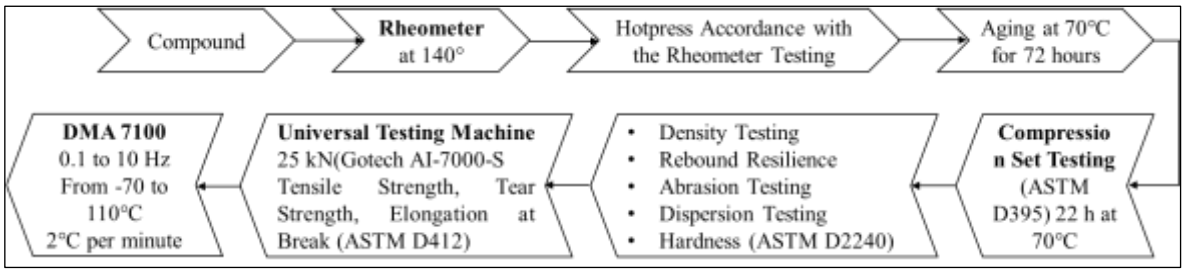


Figure 8: Step of rubber compounding testing.
Source: Authors, (2026).

III.6 TRACK-PAD PROTOTYPE FABRICATION

Based on the most optimal formulation obtained from laboratory testing, a prototype track-pad block was fabricated using a dedicated steel mold. The molding conditions followed the same curing parameters applied in laboratory vulcanization to ensure structural consistency. The prototype geometry adhered to the dimensional specifications of the original Scorpion Tank track-pad. A schematic representation of the load configuration experienced by the track pad during vehicle operation is shown in Figure 9.

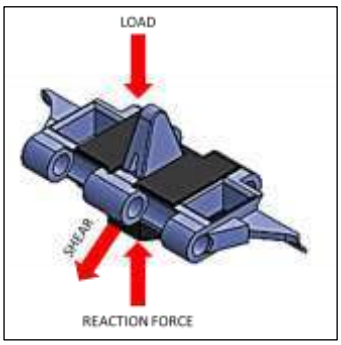


Figure 9: A schematic representation of the load configuration.
Source: Authors, (2026).

III.7 FIELD PERFORMANCE TEST

Field testing of the fabricated track pads was conducted using a Scorpion Tank belonging to the 3rd Cavalry Battalion (Yonkav 3), Malang, Indonesia. The test protocol consisted of straight-route driving for 15 km, circular turning for 10 full rotations, and a zig-zag maneuver test covering 5 km. Wear measurements were taken at intervals of 3 km to monitor progressive surface degradation, dimensional change, and mass loss. The field-test results are summarized in Table 4, and the wear-progression curve is presented in Figure 10. Data from field testing were compared with laboratory measurements to validate the reliability of the composite under real operating conditions.

IV. RESULTS AND DISCUSSIONS

IV.1 CURING CHARACTERISTICS

The curing characteristics of rubber compounds provide essential information regarding the kinetics of the vulcanization reaction. In this study, all formulations were prepared using Carbon Black N330 with different loading levels. As shown in Table 2, increasing the amount of N330 results in a gradual rise in the minimum torque (S'min), indicating an increase in the initial viscosity of the compound due to the higher solid filler content and stronger filler–rubber interactions. The maximum torque (S'max) also increases with N330 loading, although the magnitude of enhancement is not as high as typically reported for high-structure carbon blacks. This behavior aligns with the known reinforcing capability of N330, which is classified as a medium-structure grade.

The torque difference (ΔS = S'max – S'min) observed in our results shows that crosslink density improves at moderate N330 concentrations, but at higher filler loadings, ΔS begins to decline, suggesting the onset of filler–filler interactions and agglomeration that reduce the efficiency of sulfur crosslink network formation. The optimum cure time (T90) remains relatively stable across all formulations, indicating that the vulcanization process can be performed consistently without requiring major adjustment to processing conditions. Meanwhile, the scorch time (Ts2) remains within a safe processing window for all variations, ensuring that the compound retains sufficient flowability before the onset of vulcanization. Overall, the rheometer results confirm that the incorporation of N330 improves reinforcement up to an optimal level, beyond which excessive filler loading becomes detrimental to effective crosslink formation.

Table 2: Curing characteristics of track pad compounds.

Model	Curing Characteristic					
	S'max (kg·cm)	S'min (kg·cm)	ΔS (kg·cm)	t90 (min: sec)	ts2 (min: sec)	CRI (min^{-1})
Model A	30,8	4,9	25,9	8,2	2,4	16,2
Model B	32,1	5,1	27	8,1	2,25	16,8
Model C	33,4	5,45	27,95	7,55	2,1	17,5
Model D	34,2	5,7	28,5	7,4	2,05	18,1
Model E	33,9	6,05	27,85	7,55	2,15	17,6

Source: Authors, (2026).

According to Attharangsan [23], increasing filler loading reduces both scorch time and cure time, and a similar trend is observed in our predicted results for Models A–E, in which all formulations contain Carbon Black N330 at different phr levels. The difference between the maximum and minimum torque (ΔS) reflects the effective crosslink density, while the cure rate index (CRI) represents the vulcanization rate derived from the relationship between optimum cure time (T90) and scorch time (Ts2)[24]. As the phr of N330 increases, the compound generally shows lower Ts2, slightly shorter T90, and a moderate increase in ΔS up to its optimum loading. Since CRI is governed by the ratio of these parameters, the models with higher N330 phr exhibit increased CRI, except at excessive filler levels where filler–filler interactions begin to limit the vulcanization efficiency. The relationship is expressed as:

$$\mathrm{CRI} = \frac{100}{T_{90} - T_{s2}}$$

The addition of N330 caused a gradual increase in both minimum torque (S'min) and maximum torque (S'max), indicating higher initial viscosity and moderately increased crosslink density. The torque difference (ΔS) reached its highest value at a medium carbon-black loading, while excessive filler addition produced no further improvement, suggesting the onset of filler–filler interference. Both scorch time (T_{s2}) and optimum cure time (T_{90}) tended to decrease with higher phr N330, showing that the filler accelerated the vulcanization reaction. The cure rate index (CRI) varied among the models: most filled compounds exhibited lower CRI than the non-carbon-black sample due to a wider T_{90}–T_{s2} interval, whereas Model C showed the highest CRI among the filled samples, reflecting its shortest effective curing window[25]. Overall, the experimental results confirm that N330 enhances the curing rate and crosslinking efficiency at moderate levels, while excessive loading reduces curing efficiency due to increased filler aggregation.

IV.2 MECHANICAL PROPERTIES.

The rubber track pad standard in Table 3 serves as a reference for determining the required characteristic values of the rubber compound to be used in track pad applications. Several additional tests were also performed outside the standard to obtain a more comprehensive evaluation of the material's performance. The hardness of rubber vulcanizates is largely governed by the proportion of rubber to additives and the density of molecular crosslinks within the material[26]. Under an applied external load, the robustness of intermolecular bonding determines the rubber's resistance to deformation, thereby affecting its measured hardness. In this study, the hardness values of the NR–SBR/carbon black composites exhibited a clear increasing trend across all formulations, as shown in Figure 10. Compound A recorded the lowest hardness value of 83.2 Shore A, indicating that it is the softest material within the series. Compound B showed a slightly higher hardness of 84.0 Shore A, followed by Compound C at 85.6 Shore A and Compound D at 86.0 Shore A, demonstrating progressively stronger reinforcement effects.

The highest hardness was observed in Compound E, with an average Shore A hardness of 89.0, making it the stiffest and most rigid formulation among the compounds tested. Hardness is a critical parameter that reflects the stiffness, deformation resistance, and structural integrity of rubber materials. The gradual increase from 83.2 Shore A (A) to 89.0 Shore A (E) indicates that higher filler loading results in tighter molecular packing and reduced chain mobility, enhancing the compound's rigidity. These differences illustrate how variations in carbon black concentration influence mechanical performance, where softer compounds provide greater elasticity, while harder compounds exhibit better resistance to indentation and compressive deformation. Although all formulations employed carbon black N330, the variation in phr significantly affected the reinforcement behavior.

Higher carbon black content increases the effective surface area for polymer–filler interaction, resulting in stronger bonding and reduced segmental motion within the NR–SBR matrix. This explains why Compound E, with the highest filler dosage, showed the greatest hardness—approximately 5.8 Shore A, higher than Compound A. Such an increment highlights the substantial impact of filler loading on stiffness and durability. These findings are highly relevant for industrial track pad applications. Harder materials, such as Compound E, are expected to exhibit superior resistance to dynamic loading, compressive forces, and abrasive wear, thereby enhancing service life. Conversely, moderately hard compounds like A and B may provide better damping behavior but reduced long-term stability. Therefore, selecting the appropriate hardness level depends on achieving an optimal balance between flexibility, resilience, and operational durability required for track pad performance.

Table 3: Physics Testing Track Pad Standards.

Physics Testing	Rubber Standard For Track Pad
Hardness, Shore A	75
Density (gr/cm^3)	1,2
Abrasion resistance (mm^3)	350
Elongation at break (%)	228
Tensile strength (N/mm^2)	16.9

Source: [27].

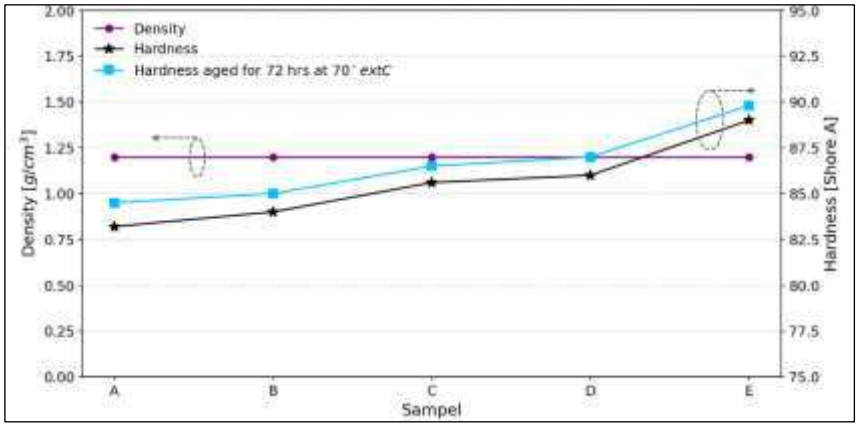


Figure 10: Density and hardness results of track pad compounds.
Source: Authors, (2026).

In the analysis of the relationship between density and hardness (as displayed in Figure 10), a clear correlation is observed for both normal and post-aging compounds. In the current dataset (Samples A through E), the Density of the material is consistently measured at 1.2 g/cm^3, indicating negligible variation in mass density across the formulations. Conversely, the Hardness (Shore A) of the normal compounds (black line) exhibits a clear upward trend, increasing from 83.2 in Sample A to 89.0 in Sample E. Since density remains constant while hardness increases, this suggests that the positive correlation is driven primarily by an increase in cross-linking density or the incorporation of stiffer filler material, rather than a change in bulk material compaction. The aging process causes the addition of cross-links and the potential for breakage of molecular chains, fundamentally altering the physical properties of the rubber.

Exposure to heat (70°C for 72 hours) induces post-curing (further vulcanization), which is evidenced by the consistently higher values of the Hardness aged compounds (blue line) compared to the normal compounds. This increase in hardness (e.g., from 83.2 to 84.5 in Sample A) confirms enhanced matrix stiffness due to the formation of additional cross-links. Understanding the aging process is vital as it can be used to predict the tensile properties of rubber and the behavior of strain energy density functions, which are critical for determining the long-term durability of the track pad material. As shown in Figure 11, the tensile strength of the NR–SBR / carbon black composites decreases progressively from Sample A (~14.8 N/mm²) to Sample E (~9.7 N/mm²), while elongation at break similarly declines from approximately 149% down to about 95%.

This trend reflects the typical behavior of carbon-black reinforced elastomers: as filler loading increases, the rubber matrix becomes stiffer and less ductile due to restriction of polymer chain mobility and formation of rigid filler networks[5]. The reduction in elongation and ultimate tensile strength at high filler concentrations can be attributed to micro-agglomeration of carbon black particles, which introduces stress concentration points and reduces the ability of the matrix to deform uniformly under load[28]. For industrial track-pad applications, these results imply a trade-off: while higher filler may improve hardness and wear resistance, excessive loading compromises tensile durability and flexibility — critical properties under dynamic and impact conditions.

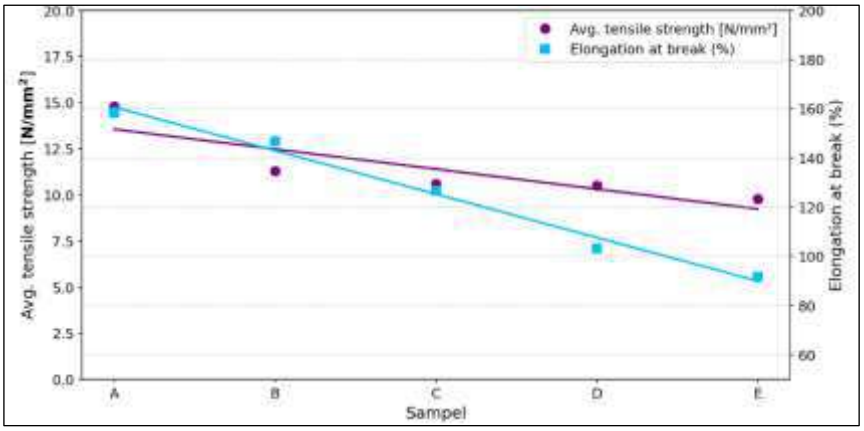


Figure 11: Tensile and elongation at break of track pad compounds.
Source: Authors, (2026).

Figure 12 presents the relationship between compression set, rebound resilience, and abrasion resistance for compounds A–E. The rebound resilience shows a clear decreasing trend, dropping from 62% (A) to 44% (E). This decline indicates reduced elasticity due to higher carbon black loading, which restricts polymer chain mobility and increases the stiffness of the vulcanized matrix. Similar findings have been reported where increased filler–polymer interactions diminish elastic energy return in NR-based composites. Conversely, the compression set values increase significantly, from 12.0% (A) to 23.8% (E).

A higher compression set signifies that the material experiences greater permanent deformation after compressive stress. This behavior is typically associated with increased crosslink density and the reduced relaxation capability of filler-rich networks, which hinder the material's ability to recover after loading. Prior studies confirm that excessive filler loading compromises elastic recovery in NR–SBR elastomers[29].

The abrasion resistance improves with higher filler content, evidenced by the decreasing abrasion volume loss from 294 mm³ (A) to 261 mm³ (E). This improvement reflects the reinforcing effect of carbon black, which enhances surface hardness and stabilizes the matrix under frictional forces. Literature widely supports carbon black's role in enhancing wear resistance due to its high surface area and strong filler–rubber bonding[30]. Overall, these results highlight a predictable performance trade-off: compounds with lower filler content (A–B) demonstrate better elasticity and recovery (high rebound, low compression set), whereas compounds with higher filler content (D–E) exhibit improved wear resistance at the cost of reduced resiliency. For track pad applications subjected to dynamic loading and abrasive ground contact, an intermediate formulation (such as Compound C) provides the most balanced mechanical profile.

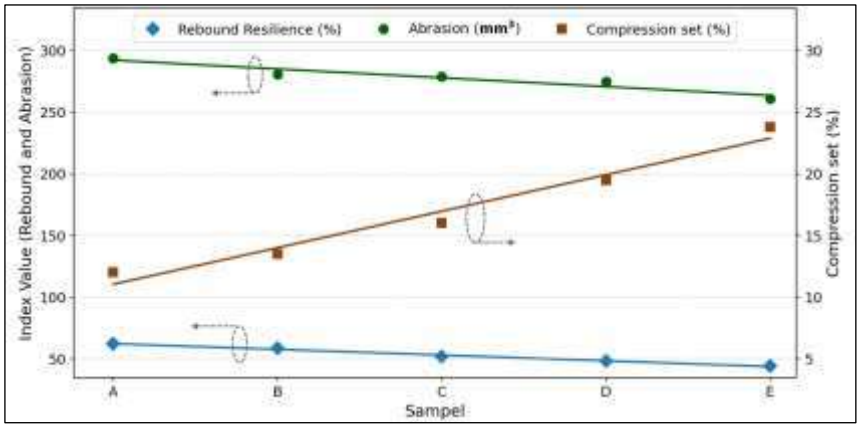


Figure 12: Rebound Resilience, compression set, and abrasion resistance of track pad compounds.
Source: Authors, (2026)

The field performance of all NR–SBR track pad formulations was evaluated through a 15-km dynamic track test conducted on an asphalt surface. Each track pad was installed on the test rig, and thickness reduction, surface wear, crack initiation, and overall structural integrity were recorded at 3-km intervals. The quantitative results are summarized in Table 4, while the wear-progression profile is illustrated in Figure 15.

Table 4: Quantitative Results of 15-km Dynamic Track Test for NR–SBR Track pads.

No	Model	3 KM	6 KM	9 KM	12 KM	15 KM
		Thickness reduction (mm)				
1	A	25.67	25.00	24.63	24.12	23.67
2	B	25.67	25.47	24.97	24.23	24.07
3	C	25.67	25.67	25.30	24.90	24.60
4	D	25.67	25.00	24.33	23.90	23.30
5	E	25.67	25.17	24.27	23.73	23.13

Source: Authors, (2026).

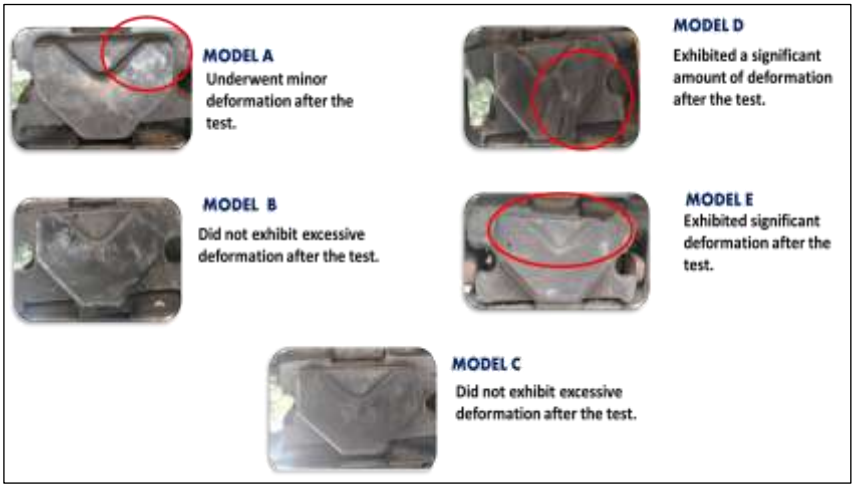


Figure 13: Wear-Progression Profile of NR–SBR Track pad Formulations During the 15-km Dynamic Track Test.
Source: Authors, (2026).

As presented in Table 4, Compounds A and B exhibited the highest rate of thickness loss due to their lower hardness and limited resistance to repetitive ground impact. Although these compounds demonstrated good elasticity, their insufficient modulus led to faster material removal and higher surface abrasion during the test. In contrast, Compounds D and E showed relatively low material loss; however, visual inspection revealed the presence of early micro-cracks and edge chipping. This failure mode is associated with excessive stiffness, which restricts energy absorption and increases localized stress concentrations under dynamic loading.

Compound C demonstrated the most balanced performance, showing moderate wear progression, minimal crack formation, and stable dimensional retention throughout the test cycle. As shown in Figure 15, the wear curve of Compound C maintained a gradual and linear trend, indicating uniform deformation behavior and superior adaptability to cyclic mechanical stresses. The qualitative damage patterns captured after the full distance, presented in Figure 13, further highlight the superior resistance of Compound C to both abrasive and impact-induced degradation. Overall, the field-test results confirm that optimal track-pad performance is not governed by a single mechanical parameter, but rather by a balanced combination of stiffness, tensile strength, abrasion resistance, and deformation capacity. The superior field durability of Compound C validates the laboratory findings and supports its suitability for real operational conditions.

IV.3 INTEGRATED PERFORMANCE EVALUATION

To determine the optimal formulation, all results—hardness, tensile strength, elongation, abrasion, and dynamic field wear—were integrated into a comparative analysis. The overall interpretation is illustrated in Figure 9, showing the relative performance of each compound. Compounds with low filler loading lack durability, while compounds with excessive filler loading are too rigid for dynamic deformation. Compound C exhibits the best balance between stiffness, strength, flexibility, and wear behavior, aligning with the performance demands of industrial track pad systems.

V. CONCLUSIONS

This study investigated the mechanical and functional characteristics of NR–SBR rubber composites reinforced with different carbon black loadings for industrial track pad applications. The experimental results confirmed that increasing filler content improves hardness and abrasion resistance but simultaneously reduces tensile strength, elongation at break, rebound resilience, and elastic recovery. Among all tested formulations, Compound C demonstrated the most balanced performance with adequate hardness, good tensile behavior, moderate compression set, and improved abrasion resistance, making it the most suitable candidate for operational track pad use.

Field testing on an armored vehicle further validated that Compound C maintained dimensional stability and exhibited minimal wear during dynamic operation. Overall, the findings indicate that an NR–SBR composite with moderate carbon black concentration offers the optimal compromise between durability and elasticity required for track pad applications. The outcomes of this study can serve as a reference for improving rubber compound formulations used in military and industrial mobility systems. Future work may include accelerated aging studies, optimization through finite element simulation, and exploration of alternative reinforcing fillers to enhance long-term performance.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Farit Hendro Wibowo and Sujito.
Methodology: Farit Hendro Wibowo.
Investigation: Farit Hendro Wibowo.
Discussion of results: Farit Hendro Wibowo, Sujito, and Mohamad Rodhi Faiz.
Writing – Original Draft: Farit Hendro Wibowo.
Writing – Review and Editing: Farit Hendro Wibowo, Sujito, and Mohamad Rodhi Faiz.
Resources: Sujito
Supervision: Farit Hendro Wibowo, Sujito, and Mohamad Rodhi Faiz.
Approval of the final text: Farit Hendro Wibowo, Sujito, and Mohamad Rodhi Faiz.

VII. ACKNOWLEDGMENTS

The authors would like to express their sincere appreciation to the promoters for their continuous support and valuable guidance in the completion of this article. This research was conducted as part of the doctoral program in the Electrical Engineering and Informatics Study Program at Universitas Negeri Malang. All authors have contributed substantially to the conception, research implementation, data analysis, and manuscript preparation, and have approved the final version of this paper for submission.

VIII. REFERENCES

[1] M. F. Omar et al., "A Comprehensive Review of Natural Rubber Composites: Properties, Compounding Aspects, and Renewable Practices with Natural Fibre Reinforcement," J Renew Mater, vol. 13, no. 3, pp. 497–538, 2025, doi: https://doi.org/10.32604/jrm.2024.057248.

[2] J. Otoranpur et al., "ANALISA BAHAN KARET NATURAL RUBBER UNTUK BOUGIE WHEEL PADA TANK SCORPION."

[3] J. Zhou, B. Tang, L. Yu, J. Song, and Z. Wang, "Application of a New Carbon Black Filler in SBR Composites," Polymers (Basel), vol. 17, no. 3, Feb. 2025, doi: 10.3390/polym17030358.

[4] S. Tang et al., "Natural Rubber/Styrene–Butadiene Rubber Blend Composites Potentially Applied in Damping Bearings," Polymers (Basel), vol. 16, no. 13, Jul. 2024, doi: 10.3390/polym16131945.

[5] C. G. Robertson and N. J. Hardman, "Nature of carbon black reinforcement of rubber: Perspective on the original polymer nanocomposite," Feb. 02, 2021, MDPI AG. doi: 10.3390/polym13040538.

[6] G. A. Espinel-Rubio and D. Feo-Ardila, "Territorio e identidad de resistencia en jóvenes del Catatumbo (Colombia), constructores de paces imperfectas," Investigación & Desarrollo, vol. 30, no. 01, pp. 40–68, Feb. 2022, doi: 10.14482/indes.30.1.303.661.

[7] M. F. Fachrizal and S. H. Suryo, "THE INFLUENCE OF MATERIAL STRENGTHON EXCAVATOR TRACK SHOE USING ABRASIVE WEAR TESTINGWITH OGOSHI HIGH SPEED UNIVERSAL WEAR TESTING METHOD," vol. 13, no. 16, 2018, [Online]. Available: www.arpnjournals.com

[8] K. Olejniczak and J. Napiórkowski, “Wear analysis of materials used for a track steering system in abrasive soil mass,” Materials, vol. 14, no. 20, Oct. 2021, doi: 10.3390/ma14206164.

[9] F. H. Wibowo, Moch. A. Choiron, and A. Purnowidodo, “Rekayasa Desain Non Pneumatic Tire dengan Struktur Hexagonal Honeycombs,” Jurnal Rekayasa Mesin, vol. 12, no. 3, pp. 701–707, Dec. 2021, doi: 10.21776/ub.jrm.2021.012.03.19.

[10] S. Tang et al., “Natural Rubber/Styrene–Butadiene Rubber Blend Composites Potentially Applied in Damping Bearings,” Polymers (Basel), vol. 16, no. 13, Jul. 2024, doi: 10.3390/polym16131945.

[11] “INTERNATIONAL ADVANCED RESEARCHES and ENGINEERING JOURNAL The effect on hardness and density of filling materials in NR/SBR rubber compounds,” 2017. [Online]. Available: www.dergipark.gov.tr

[12] N. Kinasih and D. M. I. Fathurrohman, “PENGARUH KONDISI REAKSI TERHADAP KARAKTERISTIK KETAHANAN KARET ALAM EPOKSI DALAM N-PENTANA,” 2016.

[13] A. Hasan, H. Sulistyo, and S. Honggokusumo, “VULCANIZATION KINETICS OF NATURAL RUBBER BASED ON FREE SULFUR DETERMINATION,” 2013.

[14] Nasruddin et al., “Characterization of Natural Rubber, Styrene Butadiene Rubber, and Nitrile Butadiene Rubber Monomer Blend Composites Loaded with Zinc Stearate to Be Used in the Solid Tire Industry,” Applied Sciences (Switzerland), vol. 13, no. 3, Feb. 2023, doi: 10.3390/app13031277.

[15] C. G. Robertson and N. J. Hardman, “Nature of carbon black reinforcement of rubber: Perspective on the original polymer nanocomposite,” Feb. 02, 2021, MDPI AG. doi: 10.3390/polym13040538.

[16] R. Chollakup, S. Suethao, P. Suwanruji, J. Boonyarit, and W. Smitthipong, “Mechanical properties and dissipation energy of carbon black/rubber composites,” Composites and Advanced Materials, vol. 30, Jan. 2021, doi: 10.1177/26349833211005476.

[17] K. P. Menard and N. R. Menard, “Dynamic Mechanical Analysis; Third Edition.”

[18] M. Zou et al., “Hybrid Carbon Black/Silica Reinforcing System for High-Performance Green Tread Rubber,” Polymers (Basel), vol. 16, no. 19, Oct. 2024, doi: 10.3390/polym16192762.

[19] M. T. Khan, M. Shahid, M. A. Bashir, and Q. Ulain, “Auxetic Nanocomposite based on Styrene Butadiene Rubber (SBR) Foam with Varying Nano-carbon Loading,” Manufacturing Science and Technology, vol. 3, no. 5, pp. 204–209, Dec. 2015, doi: 10.13189/mst.2015.030502.

[20] A. S. Hidayat et al., “Effect of bonding agent ratio on mechanical and physical properties of natural rubber compounds for rubber to textile adhesion applications,” Engineering and Applied Science Research, vol. 50, no. 1, pp. 82–91, 2023, doi: 10.14456/easr.2023.10.

[21] P. Sayfo, K. Pölöskei, and L. Mészáros, “Improving the mechanical and abrasion properties of silica-reinforced styrene-butadiene rubber composites by optimizing the concentrations of compatibilizers,” Polymer Bulletin, vol. 81, no. 14, pp. 12715–12731, Sep. 2024, doi: 10.1007/s00289-024-05314-w.

[22] L. A. Wisojodharmo, D. K. Arti, R. D. Pravitasari, A. S. Hidayat, A. I. Nisa, and M. N. Ariesta, “Effect of Tackifier Addition on Cushion Compound Formulation for Tire Retreading Application,” Jurnal Kimia Terapan Indonesia, vol. 21, no. 2, pp. 81–90, Apr. 2020, doi: 10.14203/jkti.v21i2.427.

[23] S. Attharangsan, H. Ismail, M. Abu Bakar, and J. Ismail, “Carbon Black (CB)/Rice Husk Powder (RHP) Hybrid Filler-Filled Natural Rubber Composites: Effect of CB/RHP Ratio on Property of the Composites,” Polymer-plastics Technology and Engineering - POLYM-PLAST TECHNOL ENG, vol. 51, pp. 655–662, Apr. 2012, doi: 10.1080/03602559.2012.662256.

[24] H. Nabil, H. Ismail, and A. R. Azura, “Compounding, mechanical and morphological properties of carbon-black-filled natural rubber/recycled ethylene-propylene-diene-monomer (NR/R-EPDM) blends,” Polym Test, vol. 32, no. 2, pp. 385–393, 2013, doi: https://doi.org/10.1016/j.polymertesting.2012.11.003.

[25] K. Ahmed, S. Sirajuddin Nizami, N. Zahid Raza, and K. Shirin, “Cure Characteristics, Mechanical and Swelling Properties of Marble Sludge Filled EPDM Modified Chloroprene Rubber Blends,” Advances in Materials Physics and Chemistry, vol. 02, no. 02, pp. 90–97, 2012, doi: 10.4236/ampc.. 2012.22016.

[26] J. Otoranpur et al., “ANALISA BAHAN KARET NATURAL RUBBER UNTUK BOUGIE WHEEL PADA TANK SCORPION.”

[27] A. A. Rachmat, T. Dirgantara, I. W. Suweca, and Y. Mardiyati, “KAJI NUMERIK GROUND PAD SHOE KENDARAAN TEMPUR DENGAN MODEL MATERIAL HYPERELASTIC,” Mesin, vol. 30, no. 1, pp. 64–74, Aug. 2024, doi: 10.5614/mesin.2024.30.1.5.

[28] V. Gopalakrishnan and C. Ramaswamy, “Patient opinion mining to analyze drug satisfaction using supervised learning,” Journal of Applied Research and Technology, vol. 15, no. 4, pp. 311–319, Aug. 2017, doi: 10.1016/j.jart.2017.02.005.

[29] “HANDBOOK OF ELASTOMERS.”

[30]Nasruddin, “Nasruddin Studi komposit Karet Alam dengan SBR dan Carbon Black dengan Fly Ash untuk Speed Bump STUDI KOMPOSIT KARET ALAM DENGAN SBR DAN CARBON BLACK DENGAN FLY ASH UNTUK SPEED BUMP COMPOSITE STUDY OF NATURAL RUBBER WITH SBR AND CARBON BLACK WITH FLY ASH FOR SPEED BUMP.”

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 774-786. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3064

ISSN ONLINE: 2447-0228

REVIEW ARTICLE OPEN ACCESS

RENEWABLE ENERGY AND RESOURCE RECOVERY SYSTEMS USING AI TOOLS

Gunasekaran Priya[1], Murugesan Kamaraj[2] and Jeyaseelan Aravind*[3]

[1, 2]Department of Biotechnology, Faculty of Science and Humanities, SRM Institute of Science and Technology, Ramapuram, Chennai 600089, Tamil Nadu, India.

[3]Department of Biotechnology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai 602105, Tamil Nadu, India.

[1]https://orcid.org/0000-0002-2971-1792, [2]https://orcid.org/0000-0002-0111-8524, [3]https://orcid.org/0000-0001-9699-2312

Email: priyag3@srmist.edu.in, drkamarajm@gmail.com, *dr.j.aravind@gmail.com

ARTICLE INFO

Article History

Received: December 6, 2025

Revised: January 10, 2026

Accepted: January 15, 2026

Published: February 28, 2026

Keywords:

Sustainable Energy Optimization

Data-Driven Energy Management

Predictive Maintenance

IoT–AI Integration

Waste-to-Energy (WTE)

ABSTRACT

The global energy transition toward sustainability demands innovative solutions that enhance efficiency, resilience, and circularity across energy and resource systems. This chapter explores the transformative role of Artificial Intelligence (AI) in optimizing renewable energy generation and resource recovery processes, highlighting the convergence of data-driven intelligence, automation, and digital infrastructure. AI techniques—ranging from machine learning and deep learning to reinforcement learning and fuzzy logic systems—are increasingly enabling predictive analytics, adaptive control, and process optimization in solar, wind, bioenergy, and hybrid energy systems. Furthermore, intelligent modeling and optimization approaches are driving progress in waste-to-energy conversion, wastewater nutrient recovery, and circular material flows, reinforcing the principles of the circular economy. The chapter presents case studies and frameworks that illustrate how AI tools enhance system performance, reduce operational costs, and support real-time decision-making. Challenges such as data quality, interpretability, and integration complexity are critically examined, along with emerging trends including digital twins, IoT–AI integration, and quantum-assisted energy analytics. Ultimately, the chapter emphasizes that the synergy between AI and sustainable resource technologies holds the potential to redefine global energy systems, paving the way toward a decarbonized and intelligent future.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The twenty-first century faces a defining challenge: ensuring the availability of clean, affordable, and sustainable energy while minimizing environmental degradation and resource depletion. The accelerating pace of industrialization, urbanization, and population growth has intensified global energy demand, putting immense pressure on conventional fossil fuel–based systems. These traditional sources contribute significantly to greenhouse gas emissions, air pollution, and climate change, necessitating a systemic transition toward renewable and circular energy models. In this context, renewable energy systems—such as solar, wind, hydro, biomass, and geothermal—have emerged as essential components of global decarbonization strategies.

However, the intermittent nature of renewable sources, the complexity of multi-resource interactions, and the increasing heterogeneity of energy networks have created new challenges in prediction, control, and optimization [1-3]. At the same time, the concept of resource recovery has gained prominence within the broader framework of the circular economy, emphasizing the reuse, recycling, and regeneration of materials, water, and energy from waste streams. Efficient recovery systems not only reduce environmental burdens but also enhance economic sustainability by converting waste into valuable inputs. Examples include waste-to-energy (WTE) conversion, anaerobic digestion for biogas production, nutrient recovery from wastewater, and smart recycling networks in industrial ecosystems.

Yet, these systems are inherently complex, governed by nonlinear processes, variable input characteristics, and evolving operational conditions—all of which demand intelligent, adaptive management solutions [4-6]. This is where Artificial Intelligence (AI)

becomes transformative. AI offers a powerful suite of computational methods capable of learning patterns from large datasets, optimizing system performance, and enabling predictive and autonomous decision-making. In the renewable energy domain, AI facilitates solar radiation forecasting, wind power prediction, demand-side management, and fault detection in grid operations. Similarly, in resource recovery, AI algorithms optimize process control, enhance biogas yield prediction, and improve waste classification and sorting accuracy. By integrating AI-driven analytics with Internet of Things (IoT) platforms, digital twins, and cloud computing infrastructures, engineers and policymakers can create intelligent energy–resource ecosystems that dynamically balance efficiency, reliability, and sustainability. The convergence of AI with renewable and resource recovery systems represents a pivotal evolution in the way energy is produced, distributed, and reused.

For instance, machine learning (ML) techniques such as regression models, decision trees, and support vector machines have been applied to optimize power output forecasting and process efficiency. Deep learning (DL) methods, including convolutional and recurrent neural networks, are enabling improved image recognition in photovoltaic (PV) inspection and enhanced time-series modeling for wind energy prediction. Reinforcement learning (RL) algorithms, inspired by behavioral psychology, are facilitating autonomous control and real-time adaptation in smart grids and bioenergy plants. Moreover, fuzzy logic systems and expert systems are being used to support complex decision-making under uncertainty—an essential capability in dynamic, multi-resource environments [7-9]. Despite its promise, the deployment of AI in renewable and resource recovery systems also presents critical challenges. Data heterogeneity, scarcity, and quality issues can limit model reliability. The “black-box” nature of many deep learning algorithms raises concerns regarding interpretability and trust, especially in safety-critical infrastructure. Additionally, the integration of AI with legacy energy systems, the need for standardized data frameworks, and cybersecurity vulnerabilities remain significant barriers to large-scale adoption.

Addressing these challenges requires interdisciplinary collaboration—bringing together energy engineers, data scientists, environmental specialists, and policymakers to design frameworks that are both technically robust and socially responsible [10-12]. The objective of this chapter is to explore the synergistic potential of AI technologies in advancing renewable energy and resource recovery systems. It provides a comprehensive overview of current methodologies, emerging tools, and practical applications where AI enhances energy conversion, recovery efficiency, and circular resource utilization. Through illustrative case studies, the chapter demonstrates how data-driven intelligence enables adaptive control, predictive maintenance, and systemic optimization across interconnected energy and resource networks. Furthermore, the discussion extends to future research directions, highlighting the role of next-generation technologies such as quantum computing, federated learning, and AI–IoT–blockchain convergence in shaping resilient, low-carbon, and self-optimizing infrastructures. As the global community strives to achieve the United Nations Sustainable Development Goals (SDGs)—particularly SDG 7 (Affordable and Clean Energy) and SDG 12 (Responsible Consumption and Production)—AI stands as a crucial enabler of the energy transition, fostering innovation that bridges technology, sustainability, and environmental stewardship [13].

II. OVERVIEW OF AI IN RENEWABLE ENERGY SYSTEMS

The transformation of the global energy landscape toward low-carbon and sustainable sources is central to achieving environmental and economic resilience. Renewable energy systems harness natural flows of energy—such as solar radiation, wind currents, hydrological cycles, and biological processes—to generate power and heat. Unlike conventional fossil fuel systems, renewable energy is virtually inexhaustible and emits minimal greenhouse gases. However, the dynamic and intermittent nature of many renewable sources presents challenges for stability, predictability, and integration into existing energy grids. In this context, Artificial Intelligence (AI) tools have emerged as indispensable enablers of efficiency, control, and optimization across renewable energy domains [14], [15].

II.1 SOLAR ENERGY SYSTEMS

Solar energy, particularly photovoltaic (PV) and concentrated solar power (CSP) technologies, is among the most mature and rapidly expanding renewable energy sources. The performance of solar systems is affected by multiple variables—solar irradiance, temperature, panel orientation, and environmental conditions. AI-based approaches are increasingly applied to model, predict, and optimize these parameters. Machine learning (ML) algorithms such as support vector regression (SVR), random forests, and gradient boosting are widely used for solar radiation forecasting, enabling more accurate power scheduling and grid balancing. Deep learning (DL) techniques, particularly recurrent neural networks (RNN) and long short-term memory (LSTM) models, have demonstrated superior capabilities in capturing the temporal dependencies of solar irradiance patterns. Furthermore, computer vision techniques based on convolutional neural networks (CNNs) are being employed for automated fault detection in PV panels, identifying issues such as soiling, shading, and microcracks from infrared and visual imagery. In addition, AI-driven maximum power point tracking (MPPT) algorithms—particularly those using reinforcement learning (RL)—enable dynamic adjustment of converter settings to maximize energy harvest under varying conditions. Through these applications, AI enhances the operational efficiency, reliability, and lifespan of solar energy systems [16-18].

II.2 WIND ENERGY SYSTEMS

Wind energy is another key component of the renewable portfolio, contributing significantly to global power generation. Yet, wind power output is inherently variable, depending on weather conditions and turbine performance. AI has proven instrumental in addressing these challenges through data-driven modeling, forecasting, and control. Artificial neural networks (ANNs) and Gaussian process models are used for wind speed and power forecasting, achieving higher accuracy than conventional statistical approaches. Predictive models assist in turbine load estimation and grid integration planning, mitigating the effects of variability. Moreover, AI-based fault detection systems employing deep learning and fuzzy inference techniques continuously monitor vibration, acoustic, and temperature data from turbines to predict mechanical failures.

This enables predictive maintenance, reducing downtime and maintenance costs. Recent advancements include the use of reinforcement learning for turbine control, where AI agents autonomously adjust blade pitch and yaw to maximize energy output under

fluctuating wind conditions. The incorporation of digital twins—virtual replicas of wind farms—further allows continuous simulation and optimization based on real-time data [19-21].

II.3 BIOENERGY AND WASTE-TO-ENERGY SYSTEMS

Bioenergy systems convert biomass, organic waste, or biogas into usable energy forms such as heat, electricity, or fuels. The heterogeneity of feedstock, coupled with nonlinear biochemical and thermochemical conversion processes, makes bioenergy management complex. AI methods are now extensively applied for process modeling, optimization, and yield prediction in anaerobic digestion, gasification, and pyrolysis systems. For example, adaptive neuro-fuzzy inference systems (ANFIS) and deep neural networks (DNNs) are employed to estimate biogas yield and optimize operational parameters such as temperature, retention time, and pH. AI models can also predict the calorific value of biomass based on its composition, supporting better feedstock blending strategies. In waste-to-energy systems, AI-assisted control algorithms dynamically regulate combustion and gasification parameters to enhance energy efficiency while minimizing emissions. By integrating AI with sensor-based monitoring and Internet of Things (IoT) platforms, bioenergy systems can operate as intelligent, self-correcting networks that continuously improve conversion efficiency and environmental performance [22-24].

II.4 HYDROPOWER, GEOTHERMAL, AND HYBRID SYSTEMS

While hydropower remains one of the most stable and established renewable sources, AI tools are increasingly being deployed to optimize reservoir management, turbine performance, and water resource forecasting. Reinforcement learning models can optimize reservoir operations by balancing flood control, irrigation needs, and electricity generation. In geothermal systems, AI techniques help in subsurface modeling, drilling optimization, and fault detection during heat extraction processes. Hybrid renewable systems—combining solar, wind, and storage technologies—are gaining attention for their ability to stabilize energy supply. AI-based multi-objective optimization algorithms, such as genetic algorithms (GA) and particle swarm optimization (PSO), are employed to determine the optimal sizing, dispatch strategy, and control parameters for hybrid microgrids. These AI-driven methods improve energy reliability while minimizing costs and emissions [25-27]. Figure 1 shows the AI applications across aenewable energy systems.

II.5 INTEGRATION CHALLENGES AND OPPORTUNITIES

Despite the demonstrated potential of AI in renewable systems, integration remains a significant challenge. Renewable energy infrastructures are distributed and data-intensive, requiring interoperable platforms that can securely collect, process, and analyze high-frequency sensor data. Moreover, ensuring data transparency, model explainability, and cybersecurity is essential for trust and adoption in energy-critical applications. However, the ongoing convergence of AI, big data analytics, IoT, and cloud-edge computing is creating new opportunities for smart, self-optimizing renewable systems. These systems not only adapt to environmental changes but also anticipate demand, predict failures, and coordinate with other resource sectors such as water and waste management—laying the foundation for a truly intelligent and sustainable energy ecosystem [28], [29].

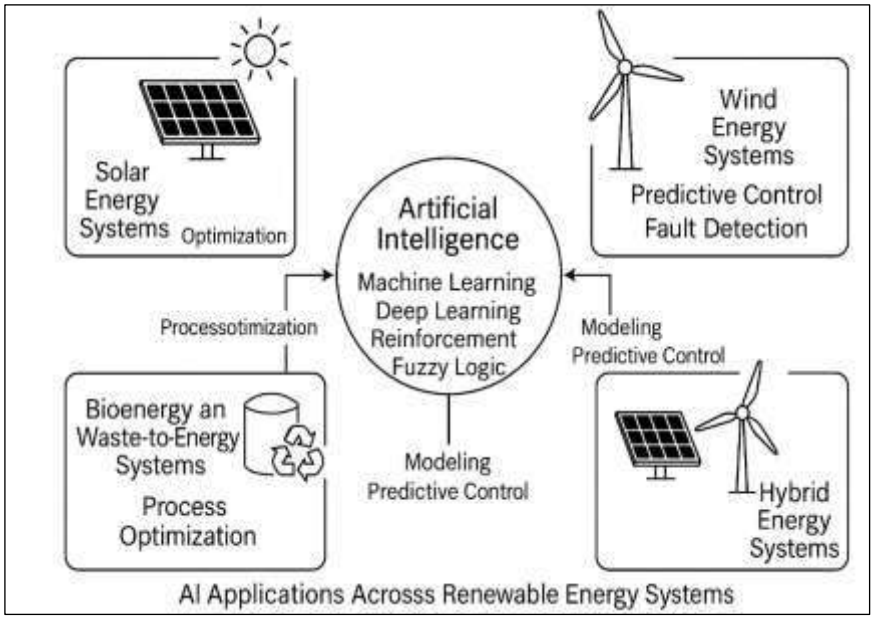


Figure 1: AI Applications Across Renewable Energy Systems.
Source: Authors, (2026).

III. RESOURCE RECOVERY SYSTEMS: CONCEPTS AND FRAMEWORKS

The growing demand for sustainable energy and material use has accelerated the shift from a traditional linear economy — based on extraction, production, consumption, and disposal — toward a circular economy (CE) model. Within this framework, resource recovery systems play a critical role in minimizing waste, conserving resources, and transforming by-products into valuable outputs.

These systems aim to extract energy, water, nutrients, and materials from waste streams using advanced technologies. The integration of Artificial Intelligence (AI) within these systems is redefining their efficiency, enabling intelligent monitoring, predictive control, and process optimization that were previously unattainable through conventional methods [30], [31].

III.1 CONCEPT OF RESOURCE RECOVERY IN THE CIRCULAR ECONOMY

Resource recovery refers to the process of reclaiming valuable components from waste materials—including solid waste, wastewater, agricultural residues, and industrial by-products—to reintroduce them into productive use cycles. The approach supports the core principles of the circular economy: reduce, reuse, recycle, and recover (4Rs). AI strengthens this concept by providing data-driven insights and automated decision-making, which allow systems to dynamically adapt to varying waste compositions, operational conditions, and demand fluctuations. Through intelligent pattern recognition, predictive analytics, and real-time control, AI-driven recovery systems optimize energy yield, minimize emissions, and support closed-loop material flows. In essence, AI acts as a cognitive layer on top of physical recovery infrastructures—transforming them into smart, self-learning systems capable of enhancing sustainability and economic viability simultaneously [32], [33].

III.2 TYPES OF RESOURCE RECOVERY SYSTEMS

Waste-to-Energy (WTE) Systems: Waste-to-energy systems convert municipal solid waste, industrial waste, or biomass into usable energy—typically in the form of heat, electricity, or fuel—via incineration, gasification, or anaerobic digestion. AI applications in WTE focus on: Process Optimization: Machine learning (ML) models predict combustion efficiency and optimize air–fuel ratios [34]. Emission Control: AI-based monitoring systems detect deviations in real time, ensuring regulatory compliance. Predictive Maintenance: Neural networks forecast component wear, reducing unplanned shutdowns. Example: Reinforcement learning algorithms have been deployed to fine-tune anaerobic digestion parameters—such as temperature and retention time—resulting in a 15–25% increase in biogas yield [35]. Wastewater Treatment and Nutrient Recovery: Wastewater contains recoverable energy and nutrients such as nitrogen, phosphorus, and organic carbon. AI-enhanced treatment plants utilize predictive modeling and automated control for: Process Control: AI regulates aeration, chemical dosing, and sludge retention time to improve energy efficiency. Fault Detection: Deep learning identifies anomalies in flow rate, pH, or dissolved oxygen levels. Nutrient Recovery Optimization: ML algorithms determine the optimal timing and method for phosphorus precipitation or ammonia stripping. Case Example: An AI-optimized membrane bioreactor (MBR) system achieved 20% lower energy consumption through adaptive control of aeration and backwashing cycles [36], [37].

Industrial Symbiosis and Circular Manufacturing: Industrial symbiosis involves the exchange of materials, energy, and waste streams among industries to minimize resource input and waste output. AI facilitates these systems by: Data Mining and Pattern Recognition: Identifying potential synergies among industries based on waste and resource profiles. Optimization Algorithms: Determining the most efficient resource exchange networks. Digital Twins: Simulating inter-industry resource flows to evaluate environmental and economic benefits. For example, in eco-industrial parks, AI-driven decision support systems have been used to optimize heat recovery networks and material reuse pathways [38]. Smart Material Recycling and Sorting Systems: Material recovery from municipal waste is often constrained by inconsistent quality and contamination. AI, especially computer vision and deep learning, is transforming recycling through: Automated Sorting: Image recognition and robotics classify materials (plastics, metals, glass, paper) in real time. Quality Control: AI systems assess purity levels of recycled streams. Lifecycle Optimization: ML tools predict recyclability potential based on material composition and degradation data. These systems enable high-throughput and precision recycling, reducing human intervention and improving material recovery efficiency [39], [40].

III.3 AI-ENABLED FRAMEWORK FOR RESOURCE RECOVERY SYSTEMS

An AI-enabled resource recovery framework typically comprises four core layers:

Sensing and Data Acquisition Layer: IoT sensors collect data on waste composition, flow rate, temperature, and emissions. Data is transmitted to centralized or edge computing platforms. Data Processing and Analytics Layer: Big data analytics and machine learning extract insights from raw sensor data. Algorithms identify inefficiencies, predict outcomes, and recommend control actions. Control and Optimization Layer: Reinforcement learning and fuzzy logic systems autonomously adjust operational parameters. Decision-support dashboards visualize performance metrics and sustainability indicators [41].

Knowledge and Decision-Making Layer: AI integrates with digital twins and optimization models for long-term planning. Insights inform policymakers and operators for adaptive resource management. This layered architecture promotes a cyber-physical integration of recovery facilities—where digital intelligence continuously enhances physical system performance [42].

III.4 SUSTAINABILITY AND ECONOMIC IMPLICATIONS

AI-driven resource recovery systems yield multiple benefits across sustainability dimensions: Environmental: Reduced greenhouse gas emissions, minimized landfill disposal, and improved water quality. Economic: Lower operational costs, enhanced energy efficiency, and monetization of recovered materials. Social: Creation of green jobs, promotion of environmental awareness, and increased resilience of urban infrastructure.

In the broader context of the energy–water–waste nexus, AI enables synergistic management of interlinked systems—facilitating sustainable urban metabolism and aligning with global sustainability goals such as SDG 6 (Clean Water and Sanitation) and SDG 12 (Responsible Consumption and Production) [43]. Examples of AI Applications in Resource Recovery Systems are listed in table 1. Figure 2 describes various AI-Driven Framework for Resource Recovery Systems.

Table 1: Examples of AI Applications in Resource Recovery Systems.

Resource Type	Recovery Process	AI Technique Used	Outcome / Benefit
Municipal Waste	Anaerobic Digestion	Reinforcement Learning	20–25% increase in biogas yield
Industrial Waste	Gasification	Neural Networks	Optimization of temperature and feed ratio
Wastewater	Membrane Bioreactor	Deep Learning	20% energy savings, predictive fouling detection
Plastic Waste	Material Sorting	Computer Vision (CNN)	95% classification accuracy in recycling lines
Eco-industrial Park	Heat Recovery	Optimization Algorithms	15% improvement in overall energy efficiency

Source: Authors, (2026).

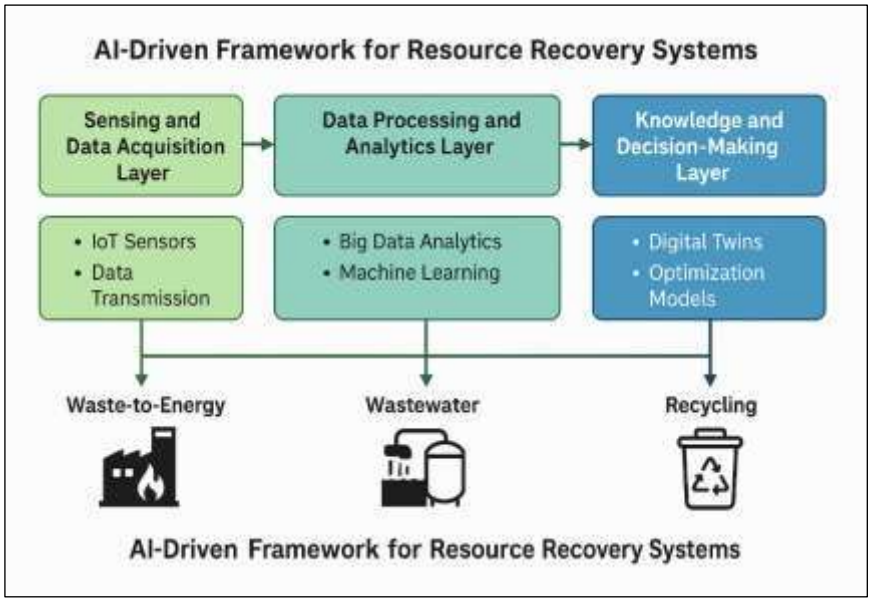


Figure 2: AI-Driven Framework for Resource Recovery Systems.
Source: Authors, (2026).

IV. ARTIFICIAL INTELLIGENCE TOOLS AND TECHNIQUES FOR RENEWABLE RESOURCE RECOVERY SYSTEMS

Artificial Intelligence (AI) encompasses a broad set of computational methodologies designed to emulate human learning, reasoning, and problem-solving. Within the domains of renewable energy and resource recovery, AI tools provide adaptive, data-driven intelligence capable of handling the nonlinearities, uncertainties, and multivariable dependencies characteristic of such complex systems. These techniques enable predictive modeling, operational optimization, and autonomous control—transforming conventional infrastructures into smart, self-learning systems that continuously evolve toward improved performance and sustainability [44].

IV.1 MACHINE LEARNING (ML)

Machine Learning (ML) is the foundation of AI applications in energy and environmental systems. It focuses on developing algorithms that can learn patterns and relationships from data without explicit programming. ML techniques are extensively used for energy forecasting, fault detection, process modeling, and optimization across renewable and recovery domains [45]. Supervised Learning: Supervised learning uses labeled datasets to train models for predictive tasks. Regression models (e.g., linear regression, support vector regression, random forest regression) are applied in solar irradiance and wind power forecasting, providing short-term generation estimates that enhance grid stability. Classification algorithms (e.g., decision trees, k-nearest neighbors, and gradient boosting) are used in waste material sorting, bioenergy feedstock classification, and equipment fault diagnosis [46].

Unsupervised Learning: Unsupervised learning identifies hidden structures within unlabeled data. Clustering algorithms, such as k-means and hierarchical clustering, support waste stream segmentation, energy consumption profiling, and anomaly detection in system operations. Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are used for dimensionality reduction and noise filtering in high-frequency energy data [47]. Applications in Resource Recovery: ML models can predict biogas yields from heterogeneous waste inputs, optimize operational setpoints in anaerobic digestion, and estimate pollutant loads in wastewater treatment plants. In material recycling, ML-based image analysis improves the accuracy of automated waste sorting systems, achieving classification accuracies exceeding 95% in some pilot studies [48].

IV.2 DEEP LEARNING (DL)

Deep Learning (DL), a subfield of ML, employs multilayered artificial neural networks (ANNs) to automatically extract features and learn complex nonlinear mappings from raw data. DL has revolutionized AI applications in energy systems, particularly in computer vision, time-series forecasting, and sensor data fusion. Convolutional Neural Networks (CNNs): CNNs are primarily used in image-based applications: Solar PV panel inspection via drone or satellite imagery to detect soiling, cracks, and degradation. Waste classification using camera-based sorting lines that distinguish materials such as plastics, paper, and metals.

Thermal imaging of turbines and reactors to detect heat anomalies or insulation failures [49]. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM): RNNs and LSTMs excel in sequential and temporal data modeling: Wind speed and power forecasting, capturing nonlinear temporal dependencies in meteorological data. Energy demand forecasting in smart grids. Process trend prediction in wastewater treatment plants, supporting proactive control [50]. Autoencoders compress and reconstruct data, useful for anomaly detection and fault diagnosis in energy systems. Generative Adversarial Networks (GANs) can simulate synthetic sensor data to augment limited datasets for training predictive models [51].

IV.3 REINFORCEMENT LEARNING (RL)

Reinforcement Learning (RL) represents a paradigm shift from traditional optimization by enabling agents to learn optimal strategies through interaction with dynamic environments. It is particularly powerful for adaptive control and decision-making under uncertainty. In renewable and resource recovery contexts: Energy Management: RL optimizes microgrid operations, balancing renewable inputs, storage, and load demand to minimize cost and emissions [52]. Process Control: In anaerobic digestion, RL agents dynamically adjust temperature, feed rate, and hydraulic retention time to maximize biogas production. Predictive Maintenance: RL-based decision systems determine the optimal timing of maintenance interventions, balancing reliability and operational costs [53]. Water–Energy–Waste Nexus: RL coordinates multi-objective management, ensuring resource efficiency across interconnected systems. The advantage of RL lies in its self-adaptive nature, making it suitable for highly variable systems where static models underperform [54].

IV.4 FUZZY LOGIC AND EXPERT SYSTEMS

While data-driven AI models rely on statistical learning, Fuzzy Logic (FL) and Expert Systems (ES) emulate human reasoning based on linguistic rules and expert knowledge. These systems are particularly effective when data are scarce, incomplete, or uncertain. Fuzzy Control: Used extensively in renewable energy plants for real-time regulation of wind turbine pitch, solar inverter voltage, and biogas plant temperature. Expert Systems: Encode domain knowledge as “if–then” rules for diagnostics, energy auditing, and decision support in resource management. Hybrid AI Systems: Combine fuzzy logic with ML/DL models (known as neuro-fuzzy systems) to merge interpretability with learning capabilities. Example: In wastewater treatment, a fuzzy control system can maintain optimal dissolved oxygen levels by adjusting aeration rates based on linguistic inputs such as “low,” “medium,” or “high.” [55].

IV.5 OPTIMIZATION ALGORITHMS AND EVOLUTIONARY COMPUTING

Optimization algorithms, inspired by natural processes, are essential in system design and operational tuning. These algorithms are often hybridized with AI models to improve convergence speed and solution accuracy. Genetic Algorithms (GA): Applied to optimize renewable energy mix, sizing of hybrid systems, and bioprocess parameters. Particle Swarm Optimization (PSO): Used for controller tuning and multi-objective energy scheduling. Ant Colony Optimization (ACO): Applied to waste transportation routing and network optimization in industrial symbiosis. By coupling AI with these algorithms, systems can achieve multi-criteria optimization (e.g., minimizing cost and emissions while maximizing efficiency) [56].

IV.6 DIGITAL TWINS AND CYBER-PHYSICAL INTEGRATION

Digital Twins (DTs) represent the convergence of AI, simulation, and real-time sensing into virtual replicas of physical systems. In renewable and recovery systems, DTs are used for system modeling, predictive maintenance, and lifecycle optimization. In wind farms, DTs simulate turbine performance and degradation patterns using ML models trained on operational data. In wastewater recovery plants, DTs predict system responses under varying loads, optimizing aeration and chemical dosing. In eco-industrial parks, DTs model interlinked material and energy exchanges to improve resource circularity. Coupled with AI and IoT, digital twins enable continuous learning, allowing operators to simulate scenarios, test control strategies, and forecast long-term performance [57]. Various AI techniques are compared in Table 2.

Table 2: Comparative Overview of AI Techniques.

AI Technique	Primary Function	Key Applications	Advantages	Limitations
Machine Learning	Pattern recognition, forecasting	Solar/wind prediction, fault detection	High accuracy, scalable	Requires quality data
Deep Learning	Feature extraction, image/time-series analysis	PV inspection, waste sorting	Handles complex data	Computationally intensive
Reinforcement Learning	Adaptive control, decision-making	Microgrid optimization, digestion control	Learns from environment	Requires large training time
Fuzzy Logic	Rule-based reasoning	Turbine control, wastewater aeration	Explainable and robust	Limited adaptability
Optimization Algorithms	System design, parameter tuning	Hybrid system sizing, resource routing	Multi-objective optimization	May converge slowly
Digital Twins	Simulation, predictive modeling	Energy plant monitoring, circular networks	Real-time adaptability	High modeling complexity

Source: Authors, (2026).

IV.7 INTEGRATION OF AI TECHNIQUES FOR HOLISTIC SYSTEM INTELLIGENCE

In practical applications, AI tools are rarely deployed in isolation. The most effective renewable and recovery systems employ hybrid AI architectures, integrating complementary methods: ML + Optimization for dynamic system design and control. DL + RL for autonomous plant management and adaptive learning. Fuzzy Logic + ML for interpretable and adaptive decision systems. Digital Twins + IoT for full cyber-physical integration and real-time feedback. Such integration leads to the emergence of self-optimizing, intelligent systems capable of adjusting operations autonomously in response to changing environmental and operational conditions [58].

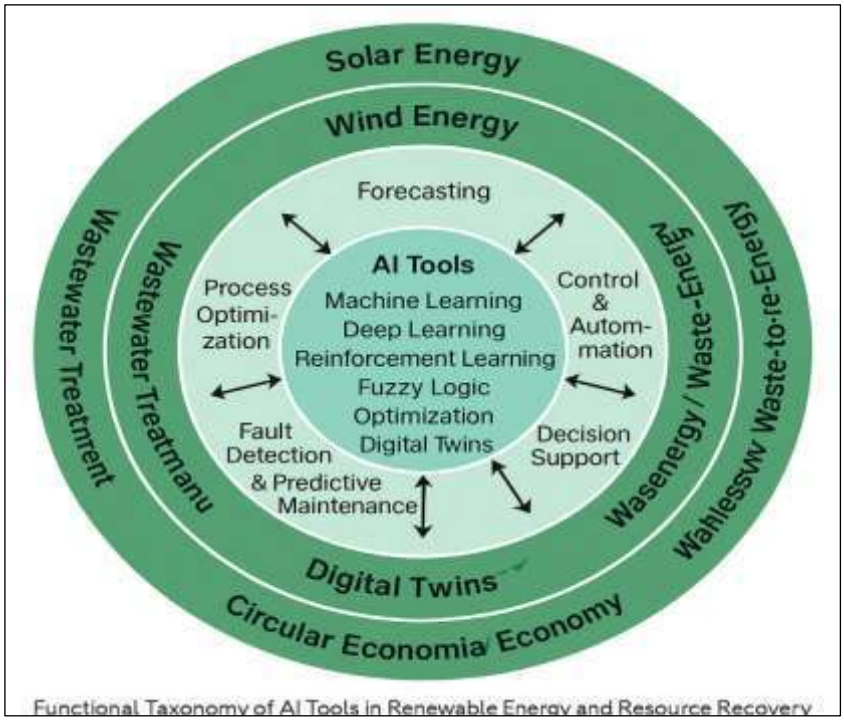


Figure 3: Functional Taxonomy of AI Tools in Renewable and Resource Recovery Systems.
Source: Authors, (2026).

V. APPLICATIONS OF AI IN RENEWABLE RESOURCE RECOVERY SYSTEM

The practical deployment of Artificial Intelligence (AI) within renewable energy and resource recovery systems has moved beyond the experimental stage into operational and industrial-scale applications. By leveraging machine learning, deep learning, reinforcement learning, and hybrid intelligence architectures, energy and environmental systems now exhibit improved reliability, adaptability, and cost-effectiveness. This section highlights notable case studies and application domains that demonstrate the integration of AI tools in renewable energy generation, waste-to-energy (WTE) plants, wastewater treatment systems, and circular resource networks [59]. The functional Taxonomy of AI Tools in Renewable and Resource Recovery Systems are portrayed in figure 3.

V.1 AI IN SOLAR ENERGY SYSTEMS

Accurate solar power forecasting is essential for grid stability, especially in regions with high photovoltaic (PV) penetration. A case study conducted by NREL (USA) employed a deep convolutional neural network (CNN) combined with satellite image data to predict solar irradiance up to 6 hours ahead, achieving 95% forecasting accuracy—a significant improvement over conventional regression models. Researchers implemented an LSTM-based model for short-term solar power prediction using real-time meteorological inputs (temperature, humidity, cloud cover). The AI system dynamically adjusted inverter settings, leading to 7–10% higher energy capture efficiency during fluctuating weather conditions [60]. At a commercial solar park in Spain, a computer vision-based deep learning system was used to inspect PV panels using drone imagery. The CNN model successfully identified cracks, hotspots, and delamination with over 96% detection accuracy, reducing maintenance downtime by 30%. Such applications demonstrate how AI contributes to predictive maintenance and extends system lifespan [61].

V.2 AI IN WIND ENERGY SYSTEMS

Wind energy operations are inherently complex due to the stochastic nature of wind and the mechanical dynamics of turbines. A collaborative project between Siemens Gamesa and Microsoft Azure AI deployed reinforcement learning (RL) algorithms to optimize wind farm control. The RL agents adjusted blade pitch and yaw settings in real time based on wind direction and load profiles, resulting in a 15% improvement in total power output while minimizing stress loads on turbine components. Similarly, deep neural networks (DNNs) trained on vibration and acoustic sensor data are now used for fault detection in gearboxes and bearings. In EU offshore wind farms, such systems have reduced unplanned maintenance by 25% and operational costs by nearly €2 million annually [62].

V.3 AI IN WASTE-TO-ENERGY (WTE) SYSTEMS

AI tools have significantly enhanced the operational efficiency of WTE facilities. In Germany, an Adaptive Neuro-Fuzzy Inference System (ANFIS) was implemented at a municipal biogas plant to optimize feedstock mixtures and retention times. The system analyzed variables such as organic loading rate, temperature, and pH to predict biogas yield in real time. The result was a 20% increase in methane production and a 10% reduction in energy input costs. A plant in Asia integrated machine learning–based predictive control with emission monitoring sensors to maintain combustion efficiency and minimize NO_x emissions. The AI controller continuously adjusted air-to-fuel ratios, leading to a 25% decrease in pollutant emissions without compromising energy output.

These examples demonstrate how AI tools transform WTE facilities into intelligent, closed-loop systems that balance energy recovery and environmental protection [63]. Wastewater treatment plants (WWTPs) are increasingly adopting AI-based automation to improve water quality, reduce operational costs, and enable nutrient recovery. In Spain's Water Reclamation Plant, a deep reinforcement learning (DRL) framework was implemented for aeration control in the activated sludge process. The AI agent learned optimal dissolved oxygen (DO) levels based on varying inflow loads and biochemical oxygen demand (BOD). The system reduced energy consumption by 18% while maintaining effluent quality within regulatory limits [64]. Table 3 list the AI Applications in Renewable Energy and Resource Recovery Systems. Figure 4 portrays AI Applications Across Renewable and Resource Recovery Domains.

V.4 AI IN MATERIAL RECYCLING AND CIRCULAR ECONOMY NETWORKS

AI is revolutionizing material recovery and recycling systems by automating sorting, improving purity, and enabling circular resource tracking. A CNN-based visual recognition system developed by TOMRA Sorting Solutions achieved 98% accuracy in material classification, distinguishing between various plastic polymers using hyperspectral imaging. The integration of AI-driven robotics in sorting lines increased throughput by 30% while maintaining quality compliance for recycled materials. At a larger scale, machine learning and digital twin models are used in eco-industrial parks (EIPs) to simulate energy and material exchanges among industries. In Spain the AI-optimized material flow modeling has improved resource reuse efficiency by 12%, demonstrating the power of AI in advancing circular economy ecosystems [65].

V.5 CROSS-SECTOR INTEGRATION: THE SMART CIRCULAR NEXUS

Recent studies emphasize that the most impactful applications of AI arise when it integrates across multiple sectors—creating smart circular nexuses that link energy, water, and waste systems. A water project developed a unified digital twin connecting wastewater treatment, biogas production, and energy recovery subsystems. Through reinforcement learning and multi-objective optimization, the integrated framework achieved a 25% reduction in total energy use and improved system-level sustainability indicators by 20–30%. These integrated systems embody the vision of autonomous, data-driven circular infrastructures, where AI serves as the coordinating intelligence managing real-time decisions and long-term sustainability objectives [66]. Role of Digital Technologies in AI-Integrated Energy and Recovery Systems are listed in table 4. Figure 5 shows the multi-layered schematic showing how AI connects with IoT, Big Data, and Cloud–Edge layers.

Table 3: AI Applications in Renewable Energy and Resource Recovery Systems.

Sector	AI Technique Used	Application	Outcome / Benefit	Reference / Region
Solar Energy	CNN, LSTM	Forecasting & Fault Detection	95% prediction accuracy, 30% maintenance reduction	USA, Spain
Wind Energy	RL, DNN	Control & Optimization	15% output gain, 25% maintenance cost reduction	Denmark
Waste-to-Energy	ANFIS, ML	Process Optimization	20% higher methane yield	Germany
Wastewater	DRL, SVM	Aeration & Membrane Maintenance	18% energy savings, 22% longer membrane life	Singapore, France
Recycling	CNN, Robotics	Material Classification	98% accuracy, 30% throughput increase	Netherlands
Circular Nexus	RL + Digital Twin	System Integration	25% energy reduction, 20% sustainability gain	EU

Source: Authors, (2026).

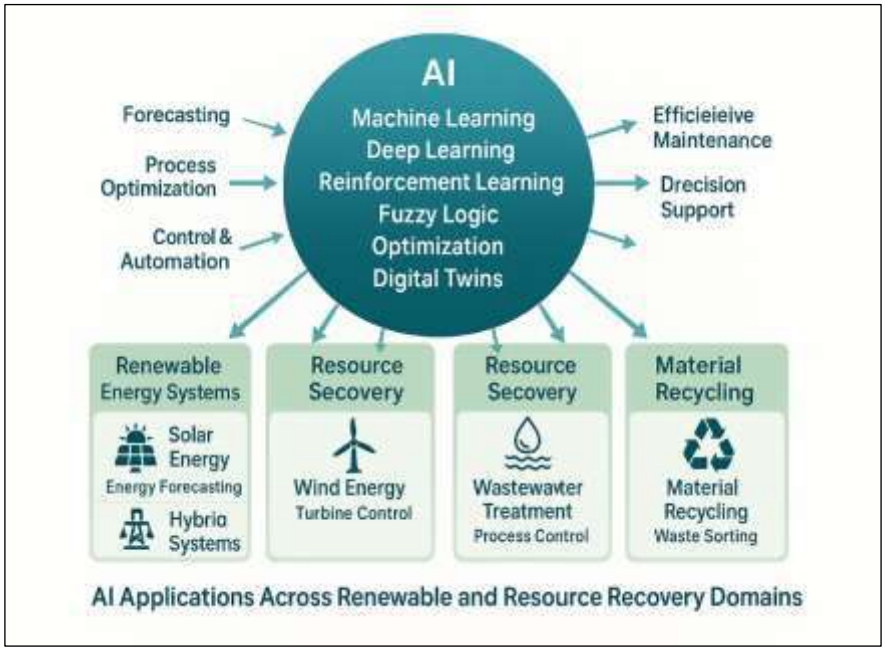


Figure 4: AI Applications Across Renewable and Resource Recovery Domains.
Source: Authors, (2026).

Table 4: Role of Digital Technologies in AI-Integrated Energy and Recovery Systems.

Technology	*Function*	*Example Application*	*Benefit*
IoT Sensors	Data acquisition	Wind turbine vibration monitoring	Real-time fault prediction
Edge Computing	Local AI inference	Solar inverter optimization	Low latency, fast response
Big Data Analytics	Pattern recognition	Wastewater quality analysis	Enhanced process control
Cloud Computing	AI model training & storage	Smart grid forecasting	Scalability & collaboration
Blockchain	Secure data exchange	Resource traceability in recycling	Transparency & trust

Source: Authors, (2026).

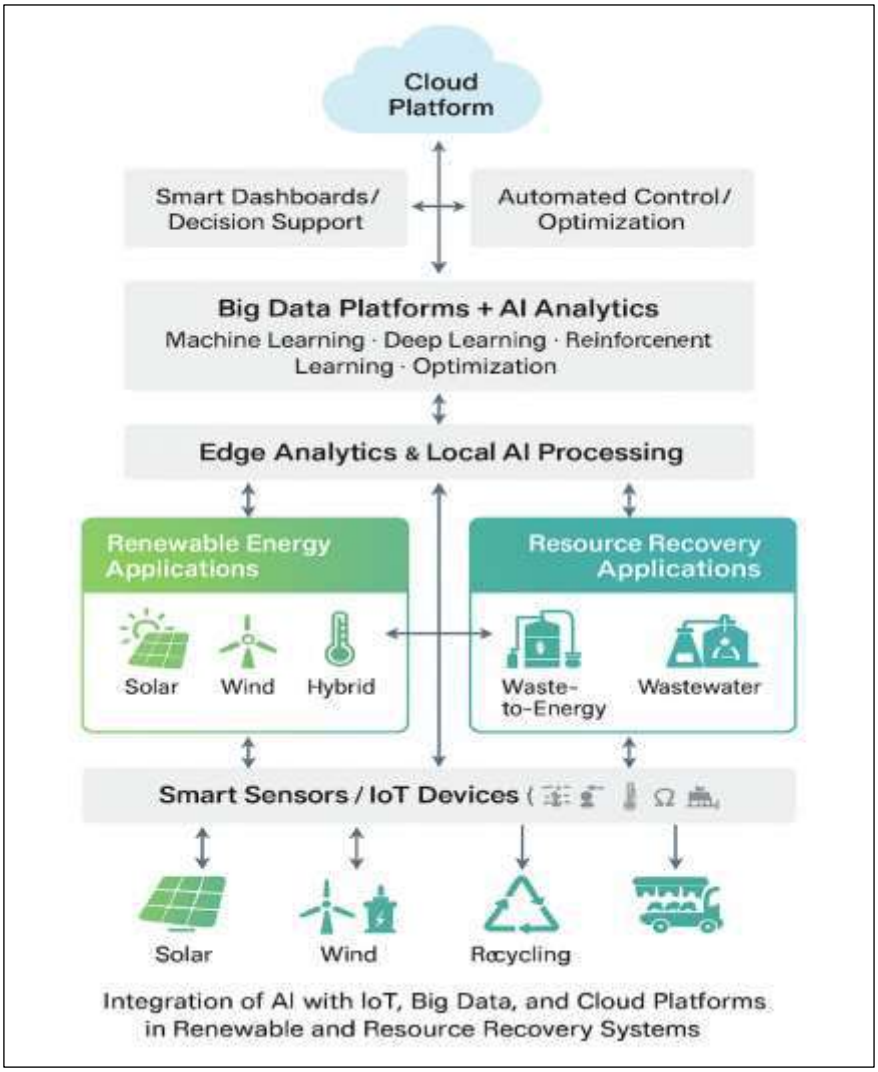


Figure 5: The multi-layered schematic showing how AI connects with IoT, Big Data, and Cloud–Edge layers.
Source: Authors, (2026).

VI. CHALLENGES AND FUTURE DIRECTIONS

While the integration of Artificial Intelligence (AI) into renewable energy and resource recovery systems has demonstrated remarkable potential, several critical challenges continue to hinder its large-scale implementation and long-term sustainability. These challenges are not merely technological—they extend across data governance, scalability, interoperability, cost-effectiveness, and ethical dimensions. Addressing these issues is essential to realize the vision of autonomous, resilient, and circular energy–resource ecosystems. The performance of AI models depends heavily on the quality, quantity, and representativeness of data. In renewable energy and recovery systems, data are often fragmented, inconsistent, or incomplete due to: Sensor malfunctions, communication latency, and missing entries. Lack of standardized data formats across different energy and waste management systems. Limited availability of labeled data for supervised learning tasks.

For example, solar and wind farms may have historical weather data, but insufficient fault or anomaly data, limiting AI's predictive capabilities. In wastewater plants, raw process data are often unstructured, making feature extraction and modeling complex. Future Direction: Developing data standardization protocols (e.g., ISO/IEC 30141 for IoT architecture) and synthetic data generation using Generative Adversarial Networks (GANs) can overcome data scarcity. Collaborative data-sharing platforms across institutions can further improve dataset diversity and reliability. The increasing connectivity of energy and resource systems exposes them to cybersecurity vulnerabilities such as unauthorized access, data tampering, or ransomware attacks.

Additionally, the collection of sensitive operational data raises concerns about privacy and data ownership—especially in industrial ecosystems involving multiple stakeholders. Future Direction: Implementing blockchain-based data management ensures immutability, traceability, and decentralized control, while federated learning frameworks allow distributed AI training without sharing raw data, enhancing both privacy and model robustness. AI models trained on site-specific data may fail when applied to different geographical, environmental, or operational contexts—a problem known as model overfitting and poor transferability. For instance, a wind turbine fault detection model developed in coastal conditions may perform poorly in high-altitude terrains due to different wind dynamics. Energy and recovery systems generate large volumes of high-velocity data, requiring real-time inference and control.

However, deep learning and optimization algorithms are computationally intensive, often demanding cloud-based GPU clusters that increase cost and latency. Future Direction: Developing lightweight AI models for Edge AI deployment can reduce dependency on cloud computation. Employing neuromorphic processors and quantum-inspired AI architectures may drastically enhance processing efficiency. Hybrid architectures combining cloud, edge, and fog computing will ensure scalability and low-latency performance. Many renewable and recovery infrastructures rely on legacy control systems (PLC/SCADA) that were not designed for interoperability with AI modules. Integrating modern AI tools requires substantial system redesign or middleware development. Future Direction: The emergence of digital twins and AI middleware platforms can bridge legacy and modern systems, enabling gradual digital transformation without full system overhauls [67].

The initial investment for AI-enabled digital infrastructures—including sensors, edge devices, and cloud services—can be prohibitive for small utilities or municipalities. Moreover, maintenance of AI systems demands specialized expertise. Future Direction: Adoption of AI-as-a-Service (AIaaS) and shared data infrastructure models can reduce financial barriers. Governments and development banks can incentivize AI adoption through green digital transition grants and carbon-credit-linked innovation funds. Regulatory frameworks governing AI deployment in energy and environmental systems remain underdeveloped. Current policies rarely address AI ethics, accountability, or algorithmic transparency, making compliance and certification challenging. Future Direction: Establish international standards for AI ethics and accountability in sustainability sectors (e.g., IEEE 7000 series). Develop AI readiness assessment frameworks for energy utilities and waste management organizations.

Promote open-source AI ecosystems to ensure transparency, reproducibility, and public trust. The deployment of AI-driven automation can raise social and ethical dilemmas, particularly regarding workforce displacement, algorithmic bias, and environmental footprint of digital infrastructure. Large-scale AI computation consumes significant electricity, potentially offsetting sustainability gains if powered by non-renewable sources. Future Direction: Emphasize Green AI principles, focusing on energy-efficient algorithms and sustainable computing infrastructure. Foster human-in-the-loop systems that blend human expertise with machine intelligence for ethical and transparent decision-making. Include AI literacy and reskilling programs in national energy transition strategies to ensure inclusive workforce adaptation.

Table 5: Summary of Key Challenges and Emerging Solutions

Challenge Area	Description	Proposed Solution / Future Trend
Data Scarcity	Limited labeled datasets	Synthetic data, federated learning
Computational Load	High energy use in AI training	Edge AI, quantum computing
Cybersecurity	Vulnerable IoT networks	Blockchain, encrypted protocols
Integration Complexity	Legacy system incompatibility	Digital twins, middleware platforms
Policy & Ethics	Lack of standards, transparency	Explainable AI, IEEE 7000 compliance

Source: Authors, (2026).

VII. CONCLUSIONS

The integration of Artificial Intelligence (AI) into renewable energy and resource recovery systems represents a transformative shift in sustainable engineering. By leveraging machine learning, deep learning, reinforcement learning, and digital twins, AI enables advanced forecasting, process optimization, and predictive maintenance across solar and wind energy, waste-to-energy, wastewater treatment, and recycling infrastructures. Supported by IoT sensing, Big Data analytics, and Cloud–Edge computing, these systems are evolving into intelligent, autonomous, and resilient cyber-physical networks that enhance energy yield, reduce emissions, and improve resource circularity in alignment with global sustainability goals. Despite these advancements, challenges such as data limitations, model robustness, legacy system integration, and cybersecurity must be addressed to fully unlock AI's potential. Emerging solutions—including federated learning, explainable AI, blockchain-based traceability, and quantum computing—offer promising avenues to overcome these barriers. As AI continues to converge with digital technologies, it will enable self-learning and self-optimizing infrastructures capable of autonomously managing energy and resource flows. Ultimately, AI is not merely a technological tool but a strategic catalyst redefining the relationship between technology, energy, and the environment. Embedding intelligence into renewable and recovery systems brings society closer to a truly sustainable, circular, and regenerative future.

VIII. AUTHOR'S CONTRIBUTION

Conceptualization: Jeyaseelan Aravind.
Methodology: Murugesan Kamaraj.
Investigation: Gunasekaran Priya.
Writing – Original Draft: Gunasekaran Priya.
Writing – Review and Editing: Murugesan Kamaraj, Jeyaseelan Aravind.
Resources: Jeyaseelan Aravind.
Supervision: Murugesan Kamaraj, Jeyaseelan Aravind.
Approval of the final text: Gunasekaran Priya, Murugesan Kamaraj, Jeyaseelan Aravind.

IX. ACKNOWLEDGMENTS

The authors thank their respective institution for all the facilities and help provided for this work.

X. REFERENCES

[1] M. J. B. Kabeyi and O. A. Olanrewaju, "Sustainable energy transition for renewable and low carbon grid electricity generation and supply," *Front. Energy Res.*, vol. 9, 743114, 2022. doi: 10.3389/fenrg.2021.743114.

[2] A. G. Olabi et al., "Renewable energy systems: Comparisons, challenges and barriers, sustainability indicators, and the contribution to UN sustainable development goals," *Int. J. Thermofluids*, vol. 20, 100498, 2023. doi: 10.1016/j.ijft.2023.100498.

[3] Q. Zeng, C. Li, and C. Magazzino, "Impact of green energy production for sustainable economic growth and green economic recovery," *Heliyon*, vol. 10, no. 17, e36643, 2024. doi: 10.1016/j.heliyon.2024.e36643.

[4] H. Elroi, G. Zbigniew, W. C. Agnieszka, and S. Piotr, "Enhancing waste resource efficiency: Circular economy for sustainability and energy conversion," *Front. Environ. Sci.*, vol. 11, 1303792, 2023. doi: 10.3389/fenvs.2023.1303792.

[5] A. K. Das, M. F. Hossain, B. U. Khan, M. M. Rahman, M. A. Z. Asad, and M. Akter, "Circular economy: A sustainable model for waste reduction and wealth creation in the textile supply chain," *SPE Polym.*, vol. 6, no. 1, e10171, 2025. doi: 10.1002/pls2.10171.

[6] M. C. Ogwu and E. A. Kosoe, "Innovative approaches to recycling, upcycling, and downcycling for sustainable waste management," *CleanMat*, vol. 2, no. 3, pp. 242–261, 2025. doi: 10.1002/clem.70013.

[7] N. E. Benti, M. D. Chaka, and A. G. Semie, "Forecasting renewable energy generation with machine learning and deep learning: Current advances and future prospects," *Sustainability*, vol. 15, no. 9, p. 7087, 2023. doi: 10.3390/su15097087.

[8] A. Nadeem et al., "AI-Driven precision in solar forecasting: Breakthroughs in machine learning and deep learning," *AIMS Geosci.*, vol. 10, no. 4, pp. 684–734, 2024. doi: 10.3934/geosci.2024035.

[9] N. F. P. Dinata, M. A. M. Ramli, M. I. Jambak, M. A. B. Sidik, and M. M. Alqahtani, "Designing an optimal microgrid control system using deep reinforcement learning: A systematic review," *Eng. Sci. Technol. An Int. J.*, vol. 51, 101651, 2024. doi: 10.1016/j.jestch.2024.101651.

[10] M. Husein, E. J. Gago, B. Hasan, and M. C. Pegalajar, "Towards energy efficiency: A comprehensive review of deep learning-based photovoltaic power forecasting strategies," *Heliyon*, vol. 10, no. 13, e33419, 2024. doi: 10.1016/j.heliyon.2024.e33419.

[11] A. V. Waghmare, V. P. Singh, T. Varshney, P. Sanjeevikumar, "A systematic review of reinforcement learning-based control for microgrids: Trends, challenges, and emerging algorithms," *Discov. Appl. Sci.*, vol. 7, no. 9, p. 939, 2025. doi: 10.1007/s42452-025-07529-6.

[12] C. J. Ejiyi et al., "Comprehensive review of artificial intelligence applications in renewable energy systems: Current implementations and emerging trends," *J. Big Data*, vol. 12, no. 1, p. 169, 2025. doi: 10.1186/s40537-025-01178-7.

[13] M. R. Shadi, H. Mirshekali, and H. R. Shaker, "Explainable artificial intelligence for energy systems maintenance: A review on concepts, current techniques, challenges, and prospects," *Renew. Sustain. Energy Rev.*, vol. 216, 115668, 2025. doi: 10.1016/j.rser.2025.115668.

[14] Q. Hassan, S. Algburi, A. Z. Sameen, H. M. Salman, and M. Jaszczur, "A review of hybrid renewable energy systems: Solar and wind-powered solutions: Challenges, opportunities, and policy implications," *Results Eng.*, vol. 20, 101621, 2023. doi: 10.1016/j.rineng.2023.101621.

[15] B. O. Eze and O. S. Ayorinde, "Prediction of renewable energy generation using machine learning a systematic review of literature," *Int. J. Innov. Res. Electron. Commun.*, vol. 11, pp. 1714–1718, 2024.

[16] W. Strielkowski, L. Civín, E. Tarkhanova, M. Tvaronavičienė, and Y. Petrenko, "Renewable energy in the sustainable development of electrical power sector: A review," *Energies*, vol. 14, no. 24, p. 8240, 2021. doi: 10.3390/en14248240.

[17] B. O. Abisoye, Y. Sun, and W. Zenghui, "A particle swarm optimization-long-short term memory (PSO-LSTM) hybrid model for forecasting global horizontal solar radiation," in *Pan-African Artificial Intelligence and Smart Systems Conference*, T. M. N. Ngatched, I. Woungang, J. R. Tapamo, S. Viriri, Eds. Cham: Springer Nature, 2025, pp. 291–308. doi: 10.1007/978-3-031-94439-0_17.

[18] O. Khouili, M. Hanine, M. Louzazni, M. A. L. Flores, E. G. Villena, and I. Ashraf, "Evaluating the impact of deep learning approaches on solar and photovoltaic power forecasting: A systematic review," *Energy Strategy Rev.*, vol. 59, 101735, 2025. doi: 10.1016/j.esr.2025.101735.

[19] K. Ukoba, K. O. Olatunji, E. Adeoye, T. C. Jen, and D. M. Madyira, "Optimizing renewable energy systems through artificial intelligence: Review and future prospects," *Energy Environ.*, vol. 35, no. 7, pp. 3833–3879, 2024. doi: 10.1177/0958305X241256293.

[20] Y. Sun and W. Han, "A review of enhancing wind power with AI: Applications, economic implications, and green innovations," *Digit. Econ. Sustain. Dev.*, vol. 3, no. 1, p. 11, 2025. doi: 10.1007/s44265-025-00059-4.

[21] K. N. Şahi N and M. Sutcu, "Probabilistic assessment of wind power plant energy potential through a copula-deep learning approach in decision trees," *Heliyon*, vol. 10, no. 7, e28270, 2024. doi: 10.1016/j.heliyon.2024.e28270.

[22] R. Garg, P. Rajput, J. Vibhandik, A. Ali, and I. Abrar, "Advances in AI-driven biomass processing: A review of conversion technologies, optimization strategies, and smart energy integration," *ACS Omega*, vol. 10, no. 42, pp. 49300–49320, 2025. doi: 10.1021/acsomega.5c05427.

[23] H. Alcocer-García, E. Sánchez-Ramírez, E. García-García, C. Ramírez-Márquez, and J. M. Ponce-Ortega, "Unlocking the potential of biomass resources: A review on sustainable process design and intensification," *Resources*, vol. 14, no. 9, p. 143, 2025. doi: 10.3390/resources14090143.

[24] N. Marzban et al., "Smart integrated biorefineries in bioeconomy: A concept toward zero-waste, emission reduction, and self-sufficient energy production," *Biofuel Res. J.*, vol. 12, no. 1, pp. 2319–2349, 2025. doi: 10.18331/BRJ2025.12.1.4.

[25] A. Aghazadeh Ardebili, M. Zappatore, A. I. H. A. Ramadan, A. Longo, A. Ficarella, "Digital Twins of smart energy systems: A systematic literature review on enablers, design, management and computational challenges," *Energy Inform.*, vol. 7, no. 1, p. 94, 2024. doi: 10.1186/s42162-024-00385-5.

[26] G. Alexakis, M. Pellegrino, L. Rodriguez-Turienzo, and M. Maniadakis, "Enhanced waste sorting technology by integrating hyperspectral and RGB imaging," *Recycling*, vol. 10, no. 5, p. 179, 2025. doi: 10.3390/recycling10050179.

[27] D. Peteleaza et al., "Water level forecasting for hydroelectric power plants using deep learning," *Renew. Energy*, vol. 256, 124692, 2026. doi: 10.1016/j.renene.2025.124692.

[28] J. V. S. Do Amaral, C. H. Dos Santos, J. A. B. Montevechi, and A. R. De Queiroz, "Energy digital twin applications: A review," *Renew. Sustain. Energy Rev.*, vol. 188, 113891, 2023. doi: 10.1016/j.rser.2023.113891.

[29] H. Gupta, P. Agarwal, K. Gupta, S. Baliarsingh, O. P. Vyas, and A. Puliafito, "FedGrid: A secure framework with federated learning for energy optimization in the smart grid," *Energies*, vol. 16, no. 24, p. 8097, 2023. doi: 10.3390/en16248097.

[30] J. C. R. Hernandez, E. Villa-Enciso, S. Cardona-Acevedo, J. Valencia, and S. Velasquez Salas, "Smart innovation for a circular economy: A systematic review of emerging trends and the future of AI in the sustainable economy," *Sustainability*, vol. 17, no. 13, p. 5793, 2025. doi: 10.3390/su17135793.

[31] A. Snoun, M. K. Mufida, A. A. El-Cadi, and T. Delot, "AI-driven innovations in waste management: Catalyzing the circular economy," *Eng. Proc.*, vol. 97, no. 1, p. 12, 2025. doi: 10.3390/engproc2025097012.

[32] D. I. Pomoni, M. K. Koukou, M. G. Vrachopoulos, and L. Vasiliadis, "Circular economy: A multilevel approach for natural resources and wastes under an agri-food perspective," *Water-Energy Nexus*, vol. 7, pp. 103–123, 2024. doi: 10.1016/j.wen.2023.12.003.

[33] E. Petelin, "Security priorities in circular economy: A conceptual review," *Sustain. Prod. Consumption*, vol. 47, pp. 655–669, 2024. doi: 10.1016/j.spc.2024.05.004.

[34] D. M. N. Bristol, I. H. V. Gue, and A. T. Ubando, "A state-of-the-art review on machine learning based municipal waste to energy system," *Cleaner Energy Syst.*, vol. 9, 100143, 2024. doi: 10.1016/j.cles.2024.100143.

[35] A. Ciobotaru, C. Corches, D. Gota, and L. Miclea, "An explainable deep learning-based predictive maintenance solution for air compressor condition monitoring," *Sensors (Basel)*, vol. 25, no. 18, 5797, 2025. doi: 10.3390/s25185797.

[36] Q. Maqsood, R. Fatima, F. Rafaqat, T. Mehmood, S. W. Ali, and M. Hussain, "Revolutionizing water and wastewater treatment: AI and machine learning-driven approaches for Advanced Solutions," *Desalin. Water Treat.*, vol. 324, 101432, 2025. doi: 10.1016/j.dwt.2025.101432.

[37] Y. Liu et al., "AI-driven solutions in wastewater treatment and agricultural reuse systems: A comprehensive review," *J. Environ. Manag.*, vol. 393, 127008, 2025. doi: 10.1016/j.jenvman.2025.127008.

[38] R. Hashmi, H. Liu, and A. Yavari, "Digital twins for enhancing efficiency and assuring safety in renewable energy systems: A systematic literature review," *Energies*, vol. 17, no. 11, p. 2456, 2024. doi: 10.3390/en17112456.

[39] M. K. Singh, S. Hait, and A. Thakur, "Hyperspectral imaging-based classification of post-consumer thermoplastics for plastics recycling using artificial neural network," *Process Saf. Environ. Prot.*, vol. 179, pp. 593–602, 2023. doi: 10.1016/j.psep.2023.09.052.

[40] S. V. T. Dao, T. M. Le, H. M. Tran, H. V. Pham, M. T. Vu, and T. Chu, "Integrating artificial intelligence for sustainable waste management: Insights from machine learning and deep learning," *Watershed Ecol. Environ.*, vol. 7, pp. 353–382, 2025. doi: 10.1016/j.wsee.2025.07.001.

[41] R. Anitha and A. Parthiban, "Smart waste ecosystems under Industry 5.0: A framework integrating digital twins, edge-AI, graph theory, and 9R circularity," *Results Eng.*, vol. 28, 107988, 2025. doi: 10.1016/j.rineng.2025.107988.

[42] R. Alsaigh, R. Mehmood, and I. Katib, "AI explainability and governance in smart energy systems: A review," *Front. Energy Res.*, vol. 11, 1071291, 2023. doi: 10.3389/fenrg.2023.1071291.

[43] E. O. Atofarati, V. O. Adogbeji, and C. C. Enweremadu, "Sustainable smart waste management solutions for rapidly urbanizing African Cities," *Util. Policy*, vol. 95, 101961, 2025. doi: 10.1016/j.jup.2025.101961.

[44] Q. Wang, Y. Li, and R. Li, "Integrating artificial intelligence in energy transition: A comprehensive review," *Energy Strategy Rev.*, vol. 57, 101600, 2025. doi: 10.1016/j.esr.2024.101600.

[45] Y. Chen, W. Gong, C. Obrecht, and F. Kuznik, "A review of machine learning techniques for building electrical energy consumption prediction," *Energy AI*, vol. 21, 100518, 2025. doi: 10.1016/j.egyai.2025.100518.

[46] V. Khandeparkar, Shreshtha, and S. K. Ramu, "Effectiveness of supervised machine learning models for electrical fault detection in solar PV systems," *Sci. Rep.*, vol. 15, no. 1, 34919, 2025. doi: 10.1038/s41598-025-18802-4.

[47] T. Dinh et al., "Data clustering: A fundamental method in data science and management". *Data Science and Management* [Preprint], vol. 18760, 2025. doi: 10.1016/j.dsm.2025.08.001.

[48] L. Chen, P. He, H. Zhang, W. Peng, J. Qiu, and F. Lü, "Applications of machine learning tools for biological treatment of organic wastes: Perspectives and challenges," *Circular Economy*, vol. 3, no. 2, 100088, 2024. doi: 10.1016/j.cec.2024.100088.

[49] A. Shaik, A. Balasundaram, L. S. Kakarla, and N. Murugan, "Deep learning-based detection and segmentation of damage in solar panels," *Automation*, vol. 5, no. 2, pp. 128–150, 2024. doi: 10.3390/automation5020009.

[50] M. Elsaraiti and A. Merabet, "Application of long-short-term-memory recurrent neural networks to forecast wind speed," *Appl. Sci.*, vol. 11, no. 5, p. 2387, 2021. doi: 10.3390/app11052387.

[51] W. Shang, J. Qiu, H. Shi, S. Wang, L. Ding, and Y. Xiao, "An efficient anomaly detection method for industrial control systems: Deep convolutional autoencoding transformer network," *Int. J. Intell. Syst.*, vol. 2024, no. 1, p. 1–18, 2024. doi: 10.1155/2024/5459452.

[52] D. Singh, O. A. Shah, and S. Arora, "Adaptive control strategies for effective integration of solar power into smart grids using reinforcement learning," *Energy Storage Sav.*, vol. 3, no. 4, pp. 327–340, 2024. doi: 10.1016/j.enss.2024.08.002.

[53] M. Marycz, I. Turowska, S. Glazik, and P. Jasiński, "Artificial intelligence in anaerobic digestion: A review of sensors, modeling approaches, and optimization strategies," *Sensors (Basel)*, vol. 25, no. 22, 6961, 2025. doi: 10.3390/s25226961.

[54] M. R. Ullah, M. Hasan, D. Biswas, M. F. Ali, and M. G. Hasan, "Techno-economic analysis of tilt angle and inter-row spacing: Optimization of a 200 MW floating solar PV plant," *Energy Convers. Manag.: X*, vol. 28, 101245, 2025. doi: 10.1016/j.ecmx.2025.101245.

[55] C. Pérez-Briceño, P. Ponce, Q. Mei, and A. R. Fayek, "A Type-2 fuzzy logic expert system for AI selection in solar photovoltaic applications based on data and literature-driven decision framework," *Processes*, vol. 13, no. 5, p. 1524, 2025. doi: 10.3390/pr13051524.

[56] B. F. Azevedo, A. M. A. C. Rocha, and A. I. Pereira, "Hybrid approaches to optimization and machine learning methods: A systematic literature review," *Mach. Learn.*, vol. 113, no. 7, pp. 4055–4097, 2024. doi: 10.1007/s10994-023-06467-x.

[57] U. Amin, D. Kim, F. N. Ahmed, G. Ahmad, and M. J. Hossain, "Digital twins for smart asset management in the energy industry: State-of-the-art," *Expert Syst. Appl.*, vol. 289, 128358, 2025. doi: 10.1016/j.eswa.2025.128358.

[58] M. Ahmadi, H. Aly, and J. Gu, "A comprehensive review of AI-driven approaches for smart grid stability and reliability," *Renew. Sustain. Energy Rev.*, vol. 226, 116424, 2026. doi: 10.1016/j.rser.2025.116424.

[59] S. Algburi et al., "The role of artificial intelligence in accelerating renewable energy adoption for global energy transformation," *unconventional resour.*, vol. 8, 100229, 2025. doi: 10.1016/j.uncres.2025.100229.

[60] Q. Paletta et al., "Advances in solar forecasting: Computer vision with deep learning," *Adv. Appl. Energy*, vol. 11, 100150, 2023. doi: 10.1016/j.adapen.2023.100150.

[61] D. Matusz-Kalász, I. Bodnár, and M. Jobbágy, "An overview of CNN-based image analysis in solar cells, photovoltaic modules, and power plants," *Appl. Sci.*, vol. 15, no. 10, p. 5511, 2025. doi: 10.3390/app15105511.

[62] P. Veers et al., "Grand challenges in the design, manufacture, and operation of future wind turbine systems," *Wind Energy Sci.*, vol. 8, no. 7, pp. 1071–1131, 2023. doi: 10.5194/wes-8-1071-2023.

[63] B. Heydari, E. A. Abdollahzadeh Sharghi, S. Rafiee, and S. S. Mohtasebi, "Use of artificial neural network and adaptive neuro-fuzzy inference system for prediction of biogas production from spearmint essential oil wastewater treatment in up-flow anaerobic sludge blanket reactor," *Fuel*, vol. 306, 121734, 2021. doi: 10.1016/j.fuel.2021.121734.

[64] O. Aponte-Rengifo, M. Francisco, R. Vilanova, P. Vega, and S. Revollar, "Intelligent control of wastewater treatment plants based on model-free deep reinforcement learning," *Processes*, vol. 11, no. 8, p. 2269, 2023. doi: 10.3390/pr11082269.

[65] O. Chacón-Albero, M. Campos-Mocholí, C. Marco-Detchart, V. Julian, J. A. Rincon, and V. Botti, "AI for sustainable recycling: Efficient model optimization for waste classification systems," *Sensors (Basel)*, vol. 25, no. 12, 3807, 2025. doi: 10.3390/s25123807.

[66] T. A. Syed, M. Y. Khan, S. Jan, S. Albouq, S. S. Alqahtany, and M. T. Naqash, "Integrating digital twins and artificial intelligence multi-modal transformers into water resource management: Overview and advanced predictive framework," *AI*, vol. 5, no. 4, pp. 1977–2017, 2024. doi: 10.3390/ai5040098.

[67] V. Varriale, A. Cammarano, F. Michelino, and M. Caputo, "Artificial intelligence in technology networks: A catalyst for achieving the SDGs," *Technovation*, vol. 151, 103398, 2026. doi: 10.1016/j.technovation.2025.103398.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 787-797. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3066

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

A LIGHTWEIGHT FEDERATED PREDICTION APPROACH FOR URBAN VRU MOVEMENT UNDERSTANDING IN AUTONOMOUS DRIVING

Lakshmi Narayana I*[1], TMN Vamsi[2]

[1]CS&SE Department, Andhra University , Visakhapatnam, Andhra Pradesh, India
[2]Department of Computer Science and Engineering, GITAM Deemed University,Visakhapatnam, Andhra Pradesh, India

[1]https://orcid.org/0000-0002-5083-6813, [2]https://orcid.org/0000-0001-6454-3934

Email: *ilnarayana1226@gmail.com, mthalata@gitam.edu

ARTICLE INFO

Article History
Received: December 6, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Autonomous vehicles;
Vulnerable road users;
Pedestrian prediction;
Cyclist motion forecasting;
Federated learning;
Social-LSTM;
YOLO detection;
SORT tracking;
Spatio-temporal modeling;
Privacy-preserving mobility systems;
Cross-domain generalization.

ABSTRACT

The growing use of autonomous vehicles in modern city transport systems shows that there is an urgent need for accurate short-term prediction of how pedestrians and cyclists will move, especially in mixed and crowded environments where movement intention, social interaction, and road layout keep changing, and this forms the main background and motivation of the study. Existing centralized learning techniques do not scale well because they face privacy rules, data ownership issues, and heavy communication requirements, which finally result in weak performance across different domains. To solve these difficulties, this research puts forward a new federated trajectory prediction approach that mixes onboard perception, lightweight tracking of detected objects, and a Social-LSTM prediction model that is improved using the FedProx algorithm, which becomes the main method contribution. The system first uses YOLO detection to find vulnerable road users, then uses SORT tracking to keep motion continuity, and trains the Social-LSTM locally while sharing only gradient updates for safe global aggregation without sharing raw sensing data. Experiments on ETH, UCY, SDD, and NuScenes datasets show reduced domain drift, better stability in different scenes, and improved ADE and FDE scores over centralized models, showing the results achieved. The concluding part states that this federated spatio-temporal learning system offers a scalable, privacy-safe, and ready-to-deploy solution for trajectory prediction, giving a new and practical step toward safer autonomous driving decisions.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

I.1 BACKGROUND AND MOTIVATION

Autonomous vehicles are becoming a regular part of modern transportation networks, and this growing presence makes it necessary to accurately predict how pedestrians and cyclists will move in short time windows, especially in busy and diverse city environments where movements are influenced by culture, road layout, and social interaction, forming the main motivation and background for this research [1], [2]. These vulnerable road users often behave in ways that do not follow fixed traffic rules, and their motion patterns shift depending on crowd density, weather, lane structures, and shared spaces, making trajectory forecasting an important requirement for safe autonomous navigation [3]. As cities become more heterogeneous and multi-modal, autonomous driving systems require prediction models that can interpret subtle cues, interpersonal interactions, and spatial constraints in real-time, and this creates a rising need for improved forecasting approaches that suit real deployment conditions [4]. Therefore, predicting VRU trajectories is not simply a sensing requirement, but a core safety factor, and it forms the foundation on which intelligent vehicle decision-making systems operate, reinforcing the need for deeper study and system development in this domain [5].

In heterogeneous metropolitan environments, variations in infrastructure design, cultural movement habits, and traffic density introduce additional challenges for motion interpretation [6]. Pedestrians may pause, accelerate, change direction, or follow implicit group

behavior, while cyclists may shift lanes, weave through vehicles, or transition between road edges and mixed pathways [4], [5]. These behaviors cannot be captured adequately by rule-based or purely geometric prediction models, requiring learning-based approaches capable of interpreting spatial context and social interaction cues [2]. As cities diversify further, autonomous systems must adapt to varied human motion patterns rather than rely on narrow training data assumptions.

The growing emphasis on safety regulations, smart mobility policy frameworks, and public acceptance further strengthens the need for dependable prediction capabilities within autonomous driving platforms 7]. Accurate VRU trajectory forecasting enhances risk assessment, supports planning modules, and reduces uncertainty in control decisions [8]. Therefore, the motivation for this research arises from the intersection of technical necessity, urban mobility complexity, human-centric safety expectations, and deployment feasibility [22]-[11]. Developing trajectory prediction models that operate effectively across diverse real-world conditions has become a foundational requirement for the progression and acceptance of autonomous mobility systems [3].

I.2 PROBLEM STATEMENT AND RESEARCH GAPS

Existing centralized learning models used for trajectory prediction are limited because they require large datasets to be collected and stored in one place, which conflicts with increasing privacy rules, ownership concerns, and regional data access restrictions, leading to major barriers in model scalability and deployment. Centralized approaches also struggle with domain shift, meaning that a model trained in one city or region performs poorly when used in another, since VRU behavior varies across cultures, infrastructure types, and mobility norms. The lack of cross-domain consistency results in unstable prediction accuracy, reduced safety margins, and unreliable behavior estimation in real road conditions, making this an open research gap that must be addressed . Furthermore, communication load, regulatory constraints, and sensor data sensitivity prevent real autonomous vehicles from sharing raw perception data, which means a new strategy is needed that allows learning without exposing private information, thereby filling an unmet requirement in current literature and real-world deployment.

I.3 NOVEL CONTRIBUTIONS

To address these unresolved challenges, this research introduces a new framework that combines federated learning with Social-LSTM trajectory prediction, using onboard sensing and lightweight tracking to enable local model learning without transmitting raw data, becoming the main contribution of the study. The proposed system uses YOLO-based detection to identify pedestrians and cyclists, applies SORT tracking to maintain consistent movement paths, and trains a localized Social-LSTM model on each vehicle, while only sending gradient updates for global aggregation under the FedProx algorithm to stabilize training across diverse environments . This combination allows the system to overcome privacy restrictions, reduce domain shift, and improve prediction robustness across different traffic scenes, making the contribution both novel and practically useful for real deployment. The work therefore offers a trajectory prediction method that is scalable, privacy-safe, and suited for heterogeneous urban environments, providing a meaningful advancement for autonomous mobility systems and intelligent driving safety .

II.RELATED WORK:

Research on trajectory prediction for intelligent transportation and autonomous mobility has increasingly focused on distributed learning strategies to overcome data access limitations, regulatory constraints, and cross-domain variability. Early work demonstrated that federated learning could enable vehicles within ad-hoc communication environments to collaboratively improve prediction accuracy without sharing raw sensor data, highlighting scalability potential in mobile networks [1]. Subsequent investigations expanded this direction by integrating secure surrounding-vehicle data into prediction models, showing that federated aggregation can enhance local inference reliability in shared roadway environments [7].

Parallel studies explored privacy-centric encryption approaches in federated prediction pipelines, indicating that homomorphic protection could preserve confidentiality while still allowing meaningful model updates, although such systems introduced computational overhead and latency concerns [10]. These foundational contributions collectively illustrate the shift from centralized architectures toward decentralized, privacy-aware predictive modeling within vehicular ecosystems. Further work addressed multi-objective optimization and adaptive coordination strategies for federated learning in traffic environments, demonstrating that vehicular networks can support distributed training while balancing accuracy, communication cost, and resource limitations [11].

Additional studies proposed destination-guided LSTM architectures operating under federated settings, showing that embedding intention cues could improve pedestrian trajectory forecasting in shared urban spaces [6]. Stability challenges in distributed training were also examined, where researchers identified gradient divergence, node variability, and asynchronous participation as key factors affecting model consistency [12]. Broader surveys and analytical studies emphasized that federated learning remains a developing paradigm within intelligent transportation systems, noting unresolved issues regarding domain heterogeneity, communication bottlenecks, and deployment feasibility in mixed mobility landscapes [2].

These findings indicate that although federated frameworks offer promising advantages, practical and methodological challenges persist. Reviews of federated methods for vehicular networks further reinforced the importance of privacy-preserving collaboration and cross-regional data protection, summarizing that distributed learning approaches are increasingly relevant for emerging automotive communication infrastructures [13]. Research addressing automotive-specific deployment scenarios demonstrated that federated learning could be integrated into real vehicle platforms, although adaptation, synchronization, and inference reliability were identified as limiting factors [14]. Complementary studies introduced secure proxy reencryption mechanisms to strengthen privacy layers within vehicular federated systems, highlighting the ongoing tension between security enforcement and computational feasibility [8].

Together, these works reveal that while federated approaches are advancing, existing frameworks often prioritize communication protocols or cryptographic integration rather than motion prediction accuracy for vulnerable road users. Parallel to federated learning developments, researchers in trajectory prediction explored scene-aware models capable of interpreting lane structure information and

imitative driving policy guidance, demonstrating improvements in forecasting precision for autonomous navigation tasks [15]. Complementary studies analyzed spatio-temporal learning models designed for full self-driving systems, showing that integrating temporal dependencies with environmental interpretation could support more reliable prediction outcomes in dynamic roadway settings [3]. However, these approaches typically relied on centralized datasets and lacked mechanisms for cross-domain adaptability, limiting their suitability for real-world deployments where urban environments vary significantly. Moreover, many existing models focused on vehicle trajectory prediction and did not address pedestrian and cyclist behaviors, which represent more irregular and socially influenced movement patterns.

II.1 PERCEPTION AND DETECTION MODULE (YOLOV11)

The perception stage in the proposed framework employs the YOLOv11 detection model because it provides fast inference speed, high sensitivity toward vulnerable road users, and suitability for real-time deployment on autonomous vehicle hardware. YOLOv11 processes the input frame I∈RH×W×3 and predicts object locations using a single forward pass, reducing latency compared to multi-stage detectors[16]. The model operates using an input resolution of 640×640, which balances spatial detail and computational load, allowing detection of pedestrians and cyclists even under partial occlusion or motion blur[17]. The prediction head estimates bounding boxes B=(x,y,w,h), objectness score so, and class confidence sc, producing final confidence S=so×sc This anchor-free formulation removes reliance on predefined anchor box templates, allowing more flexible detection in dense urban scenes where VRU shapes and scales vary strongly. The convolutional backbone extracts hierarchical features, and feature pyramid fusion enhances detection of small targets, which is necessary since pedestrians and cyclists often occupy limited pixel regions in wide-angle automotive camera feeds.

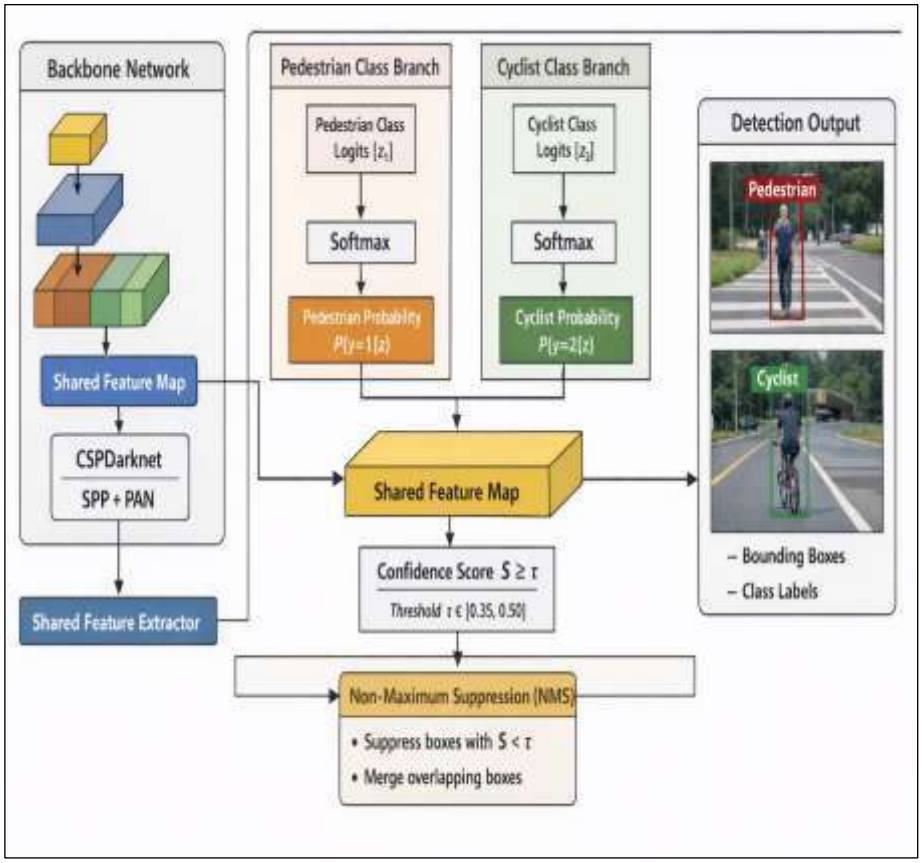


Figure 1: Yolov11 architecture.
Source: [3].

YOLOv11 provides multi-class VRU handling by applying a shared feature representation and separate class logits for pedestrians and cyclists, enabling the model to distinguish similar movement silhouettes. The classification layer outputs probability distributions using the softmax function

$$P(y = k \mid z) = exp(z_k)/(\sum_(j = 1)\^K exp(z_j)), \quad k \in \{1,2\}. \tag{1}$$

Where K=2 for VRU categories. Confidence thresholding is applied with a decision rule S≥τ, where τ is typically set between 0.35 and 0.50 to reduce false positives without suppressing true detections in crowded environments. This step stabilizes detection outputs and produces reliable identity candidates for downstream tracking. The model thereby supplies clean, continuous detection streams that support accurate trajectory construction, forming a perception layer suited for real-world mixed-traffic navigation.

II.2 MULTI-OBJECT TRACKING MODULE (SORT)

The tracking stage employs the SORT algorithm to maintain consistent identities of pedestrians and cyclists across consecutive frames, using a Kalman filter to predict future bounding box states and correct them based on new detections [18]. The filter models each tracked VRU with a state vector representing position and velocity, enabling short-term motion estimation even when temporary occlusions occur.

$$P_(t|t-1) = FP_(t-1|t-1)\, F\hat{}T + Q.. \tag{2}$$

After (2)prediction, the Hungarian assignment algorithm performs optimal matching between predicted tracks and YOLOv11 detection outputs using an IoU-based cost matrix, ensuring that each detection is paired with the most probable track.

Track ID persistence is maintained by incrementing or retaining identifiers based on confirmed matches, while unmatched tracks enter a temporary retention state to avoid premature deletion. Noise handling is achieved through covariance updates, allowing uncertainty to expand when measurements are missing and contract when reliable detections are available. This produces stable trajectory continuity and reduces fragmentation in dense or cluttered urban scenes, supplying reliable sequential coordinates for Social-LSTM processing.

II.3 TRAJECTORY SEQUENCE CONSTRUCTION

The trajectory construction stage converts the SORT-generated bounding boxes into coordinate paths by extracting each VRU's centroid position (xc,yc) per frame and organizing them into time-ordered sequences for prediction processing[19]. When intermittent detection gaps occur, interpolation is applied to estimate missing points, while smoothing filters reduce jitter caused by sensor noise or rapid bounding box fluctuations. A social interaction neighborhood radius is defined to determine nearby agents whose movements influence the target trajectory, allowing the system to model local group dynamics and collision-avoidance behavior. The resulting refined sequences form consistent spatio-temporal inputs for the Social-LSTM prediction module.

II.4 SOCIAL-LSTM PREDICTION MODEL

The Social-LSTM module predicts future positions of pedestrians and cyclists by modeling their temporal motion patterns using recurrent sequence encoding and hidden state propagation [15], [20]. Each VRU trajectory sequence (xt,yt) is fed into an LSTM network that captures motion dynamics through gated memory operations, allowing the model to retain directional intent and velocity evolution across time . The hidden state ht and cell state ct represent internal motion context and temporal continuity, enabling the network to anticipate future displacement trends as shown in equation 3.

$$\begin{aligned} 〚\quad i〛_t &= \sigma(W_i\,[x_t, h_(t-1)\,] + b_i\,) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ h_t &= o_t \odot tanh(c_t\,) \end{aligned} \tag{3}$$

A social pooling mechanism aggregates neighboring VRU states within a spatial grid, allowing the architecture to incorporate interpersonal influences such as collision avoidance, path yielding, and lane negotiation. This grid-based pooling resolves interaction dependencies by integrating the hidden states of nearby agents into the target agent's predictive context, allowing more realistic movement forecasting in dense metropolitan scenes [21], [22].The interaction modeling component extends beyond isolated trajectory encoding by fusing spatial proximity and temporal sequencing, enabling the Social-LSTM to infer implicit group behaviors, directional shifts, and shared flow patterns [23], [24]. Predicted future coordinates are generated through a decoding layer that maps the hidden representation into displacement vectors or global positions, capturing motion continuity and uncertainty. The model outputs multi-step predictions across a fixed horizon, supporting downstream navigation and risk assessment.

$$H_t\hat{}pool = \sum_(j \in N_i)\, h_t\hat{}j \tag{4}$$

By leveraging pooled interaction features and recurrent state evolution, the Social-LSTM architecture provides more adaptive and human-like forecasting performance compared to purely geometric or transformer-free temporal models, forming a suitable predictive core for the proposed federated deployment setting [25], [26].

II.5 FEDPROX-BASED AGGREGATION

The federated learning framework enables each autonomous vehicle to perform a local training cycle using only its onboard perception-derived VRU trajectory data, ensuring that raw sensor streams never leave the device [27]. Local model updates are computed after multiple mini-batch iterations and then transmitted during communication rounds to a coordinating server, where model update packaging reduces payload size by sharing only gradient differentials instead of full parameter sets. The no–data-sharing principle ensures privacy compliance, regional data governance compatibility, and operational safety by preventing exposure of identifiable VRU behavior patterns [28], [12]

$$(min)_w\; F_k\,(w) + \mu/2 \parallel w - w\hat{}((t)\,) \parallel \hat{}2 \tag{5}$$

Optional features such as device participation rate allow partial involvement when vehicles are offline, while straggler handling enables asynchronous aggregation so that slower devices do not halt global progress. Communication compression techniques, including quantization and sparsification, further lower bandwidth requirements for large-scale deployment in dense mobility networks [2], [29].

This distributed structure supports scalable training across heterogeneous traffic regions while maintaining embedded inference feasibility for onboard prediction modules.

Table 1: Comparison of Learning Approaches.

Approach	Data Handling	Privacy Level	Accuracy	Security
Centralized Learning	Aggregated	Low	Variable	Weak
FedAvg	Local Updates	Medium	Moderate	Basic
FedProx	Local Constrained	High	Stable	Enhanced

Source: Authors, (2026).

The FedProx aggregation strategy is used instead of the traditional FedAvg method because it offers improved stability under heterogeneous and non-IID VRU trajectory distributions collected from varied metropolitan environments [8].

FedProx introduces a proximal term into the optimization objective, reducing divergence between local model updates and the global reference model, which is crucial when pedestrian and cyclist behaviors differ across infrastructure layouts and cultural motion norms [30]. In comparison, FedAvg assumes homogeneous data distributions and suffers from convergence degradation when local motion patterns vary significantly.

$$w^{(t+1)} = \sum_{k=1}^{K} \quad \frac{n_k}{n} w_k^{(t)} \tag{6}$$

The proximal constraint enhances training robustness by limiting gradient drift and ensuring smoother global update convergence, even when participating vehicles contribute unevenly sized datasets or inconsistent update frequencies. Under non-IID VRU data, FedProx therefore provides faster stabilization, reduced oscillation, and higher predictive coherence in cross-domain forecasting environments [15], [22].

III. PROPOSAL METHOD

The proposed methodology introduces an integrated perception-to-prediction pipeline designed to operate within autonomous vehicle platforms while preserving data privacy and supporting heterogeneous urban environments. The system begins with onboard sensing, where YOLOv11 processes each incoming frame to detect pedestrians and cyclists, producing bounding boxes and class scores that form the perception foundation. These detections are refined through non-maximum suppression and confidence filtering to ensure reliability under motion blur, occlusion, and dense metropolitan traffic. The output of this stage provides structured VRU observations that are used to build temporal paths, forming the basis for downstream trajectory forecasting in real-world mobility conditions. This combination of lightweight inference and real-time processing makes the perception layer compatible with embedded automotive hardware and aligns with deployment requirements in autonomous navigation contexts.

Following detection, the SORT tracking mechanism associates VRU positions across consecutive frames and assigns persistent identifiers to maintain continuous motion histories. The Kalman filter predicts the next bounding box state using a linear motion estimate, while the Hungarian algorithm matches detections to predicted tracks based on minimum cost assignment. Resulting trajectories are converted into centroid coordinate sequences, and interpolation is applied to fill short detection gaps caused by occlusion. A neighborhood interaction radius defines nearby agents whose motion influences the target VRU, enabling the construction of socially contextualized temporal sequences. These spatio-temporal paths serve as input to the Social-LSTM model, which performs multi-step motion forecasting based on historical movement behavior and localized group dynamics.

The Social-LSTM prediction module employs recurrent temporal encoding to represent movement evolution through hidden state propagation. The LSTM gate operations retain directional continuity, while social pooling aggregates the hidden states of neighboring VRUs into a shared interaction tensor that reflects interpersonal influence. The hidden representation ht is then transformed by a decoding layer to produce predicted future displacement vectors across the forecasting horizon. This architecture models motion uncertainty, behavior intention, and collision-avoidance tendencies, offering improved realism compared to purely geometric approaches. The prediction output supports autonomous planning modules, enabling safer navigation and reduced trajectory ambiguity in dynamic environments.

To enable distributed learning without exposing sensitive VRU movement data, the methodology incorporates a federated learning framework, where each autonomous vehicle performs localized model training on its own trajectory sequences. At communication round rrr, each participating device updates its model parameters wk(r) for a fixed number of local epochs and transmits only the resulting gradients or weight deltas to the central aggregator. No raw perception data, images, or coordinates leave the vehicle, ensuring full compliance with privacy and governance policies. Device participation rate, straggler tolerance, and gradient compression mechanisms further optimize communication efficiency, allowing scalable deployment in connected transportation networks.

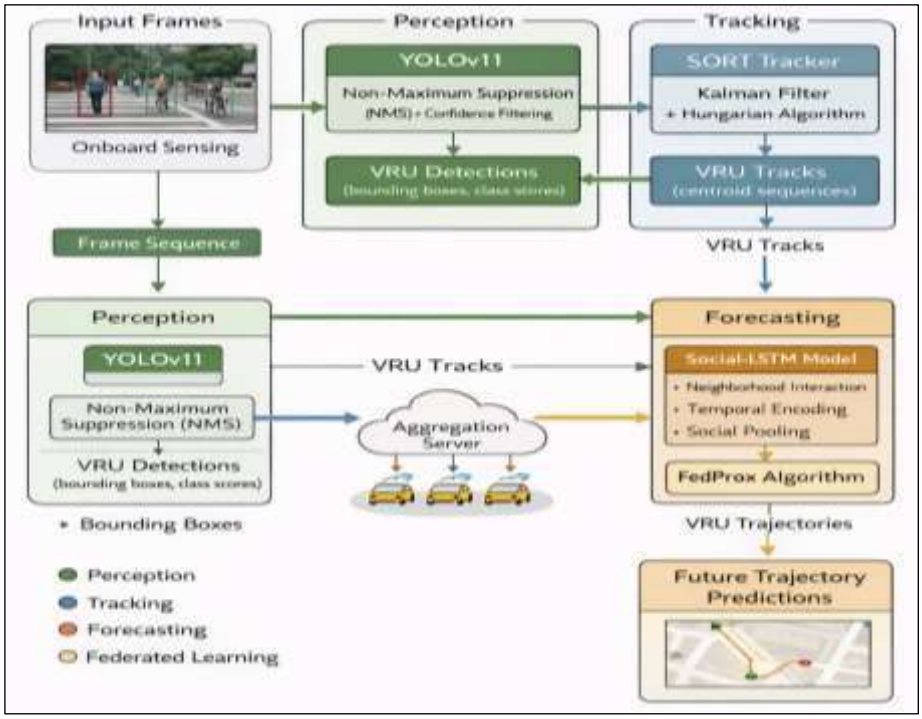


Figure 2: Proposal Method.
Source: Authors, (2026).

The aggregation server applies the FedProx algorithm instead of standard FedAvg to stabilize convergence under non-IID trajectory distributions caused by differing urban layouts, cultural movement patterns, and VRU density variations. FedProx introduces a proximal regularization term into the optimization objective, constraining each local update to remain closer to the global parameter state and reducing divergence across regions. The modified objective is expressed as:

$$\min_{w} F_k(w) + \frac{\mu}{2} \| w - w^{(r)} \|^2 \tag{7}$$

Where μ controls the proximity strength. This improves cross-domain generalization and yields more consistent prediction accuracy across diverse deployment scenarios. The resulting federated Social-LSTM model forms a scalable, privacy-preserving, and domain-robust trajectory forecasting solution suitable for real-world autonomous mobility systems.

IV.DATA PRESENTATION:

IV.1 DATASET COLLECTION:

The dataset collection process integrates four benchmark trajectory prediction datasets ETH, UCY, SDD, and NuScenes to ensure diversity in pedestrian and cyclist motion patterns across different urban scenarios[10]. The ETH and UCY datasets contribute approximately 25,000 annotated frames capturing social crowd interactions in open pedestrian zones, while the SDD dataset provides nearly 60,000 high-resolution frames featuring mixed-traffic VRU behavior at campus intersections. NuScenes adds around 40,000 multi-modal frames containing dense metropolitan traffic, cyclists, and sidewalk users recorded from autonomous vehicle sensors[8]. All datasets were standardized to a unified sampling rate, cleaned for missing annotations, and structured into observation prediction sequences for model training and evaluation.



Figure 3: Dataset Collection.
Source: Authors, (2026).

IV.2 DATASET PREPARATION AND PREPROCESSING

The dataset preparation phase focused on harmonizing four heterogeneous datasets ETH, UCY, SDD, and NuScenes into a unified format suitable for federated trajectory learning. Each dataset originally differed in frame rate, coordinate system, annotation format, and VRU density[3], [30]. Therefore, all video sequences were standardized to a uniform sampling rate of 2.5–3 Hz to balance temporal resolution and computational overhead. Pedestrian and cyclist annotations were transformed into consistent world-coordinate representations using homography-based mapping for ground-plane projection. Noise in bounding box coordinates was mitigated using temporal smoothing filters, while missing trajectory points were reconstructed via linear interpolation to maintain sequence continuity. A sliding window approach generated fixed-length observation segments Tobs and prediction targets Tpred ensuring uniformity in sequence structure across datasets.

Additional preprocessing steps addressed domain heterogeneity and scene variability to support robust Social-LSTM learning. Each trajectory was normalized by subtracting the initial reference point to remove global positional shifts, and velocities were computed using first-order temporal differences to embed motion dynamics. Social interaction neighborhoods were defined using a fixed-radius spatial grid to identify VRU neighbors influencing the target agent's motion, ensuring compatibility with social pooling operations. Scene-level data augmentation including trajectory mirroring, temporal stretching, and minor spatial perturbations was applied cautiously to enhance generalization while preserving behavioral realism. All cleaned and processed sequences were partitioned into training, validation, and testing sets following non-overlapping scene splits to enable unbiased cross-domain evaluation in a federated learning environment.

IV.3 EXPLORATORY DATA ANALYSIS (EDA)

The EDA focused on examining label distributions, spatial–temporal patterns, and inter-variable relationships across all datasets. A correlogram of trajectory-derived labels position (x,y) velocity (vx,vy), acceleration (ax,ay), and social distance dij was generated to analyze linear dependencies using Pearson correlation

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{8}$$

Strong correlations were observed between velocity and displacement magnitude, indicating consistent motion continuity, while acceleration showed higher variance under dense interactions. Label frequency plots revealed class imbalance between pedestrians and cyclists, requiring normalization during training. Spatial heatmaps highlighted congestion zones, and temporal EDA confirmed variable crowd density across scenes, validating the need for federated, domain-aware prediction.

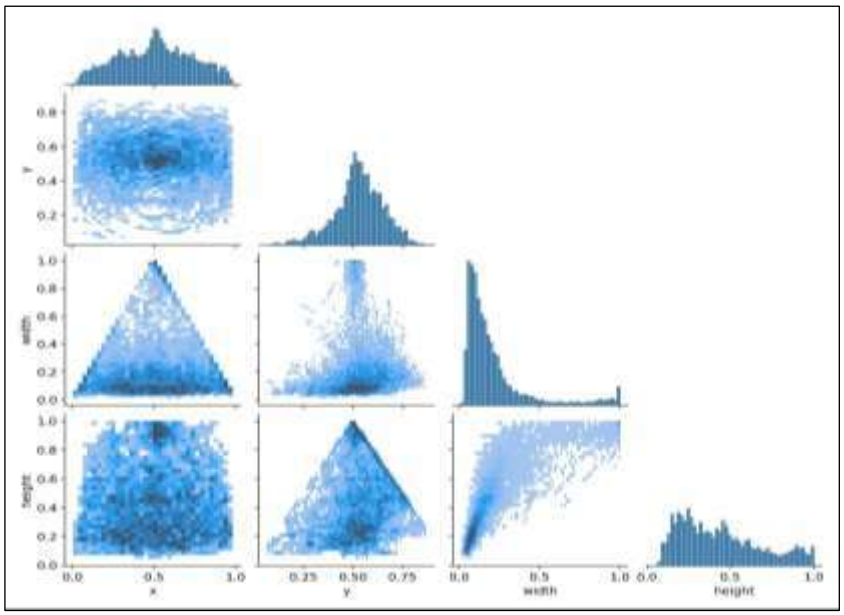


Figure 4: EDA.
Source: Authors, (2026).

V.IMPLEMENTATION

V.1 PERCEPTION AND TRACKING IMPLEMENTATION

The implementation begins with the YOLOv11-based perception pipeline, where each input frame $I \in R^{H \times W \times 3}$ is resized to a standardized resolution of 640×640 to ensure stable detection throughput across all datasets. YOLOv11's anchor-free detection head generates bounding-box predictions (x,y,w,h), objectness score so, and class probability sc. The final detection confidence is computed using

$$S = so \times sc \tag{9},$$

And detections below a confidence threshold τ=0.4 are discarded. Class-wise VRU detection (pedestrians, cyclists) enables reliable segmentation of mixed-traffic participants in dynamic urban environments.

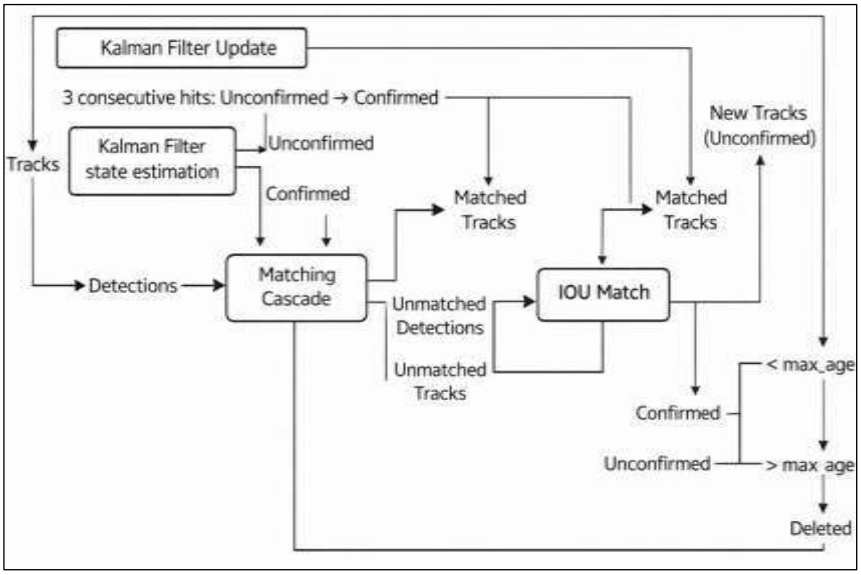


Figure 5: Workflow showing SORT tracking with Kalman updates, matching, and track management.
Source: Authors, (2026).

To maintain continuous identities across frames, the SORT tracker applies frame-to-frame association. A Kalman filter predicts the next bounding-box state using a linear motion model, while the Hungarian algorithm matches prediction-to-detection pairs using an IoU-based cost matrix

$$Cost(A,B) = 1 - IoU(A,B) \tag{10}$$

Unmatched tracks are temporarily retained before termination to avoid premature trajectory fragmentation. This ensures stable identity tracking, enabling consistent time-indexed trajectories for subsequent LSTM-based modeling.

V.2 TRAJECTORY CONSTRUCTION AND SOCIAL MODELING

Bounding box coordinates are converted into trajectory centroids (xc,yc) forming temporal sequences. Missing points due to detection gaps are reconstructed using linear interpolation:

$$x_t = \alpha x_{t-1} + (1-\alpha)x_{t+1}\; y_t = \alpha y_{t-1} + (1-\alpha)y_{t+1} \tag{11}$$

Where 0<α<1. A smoothing filter is applied to reduce jitter caused by sensor noise. Each VRU trajectory is segmented into observation windows Tobs and prediction horizons Tpred, ensuring uniform temporal structure across all datasets.To embed interpersonal influence, each VRU is assigned a spatial neighborhood grid. Agents within a radius RRR are considered socially relevant, computed as:

$$N_i = \{j \mid\mid\mid \left(x_t^i, y_t^i\right) - \left(x_t^j, y_t^j\right) \|_2 \leq R\} \tag{12}$$

The hidden states of these neighbors are aggregated through social pooling, enabling the system to incorporate collision avoidance, cooperative movement, and group dynamics into the predictive model.

V.3 FEDERATED LEARNING AND FEDPROX AGGREGATION

Each autonomous vehicle performs local training of the Social-LSTM model on its own VRU trajectories. During communication round r, each client updates parameters wk(r) for EEE epochs and sends only the model deltas Δwk to the server. No raw trajectory or sensor data is transmitted, ensuring compliance with privacy policies. The global update under FedAvg is traditionally computed as

$$w^{(r+1)} = \sum_{k=1}^{K} \frac{n_k}{n} w_k^{(r)} \tag{13}$$

However, non-IID VRU distributions make FedAvg unstable across heterogeneous cities.FedProx addresses this instability by introducing a proximal term that constrains local updates, minimizing drift across domains. The modified optimization objective as shown in equation(5) .This prevents excessive divergence caused by region-specific movement patterns. Aggregated updates yield a domain-robust global model capable of consistent performance across varied metropolitan environments.

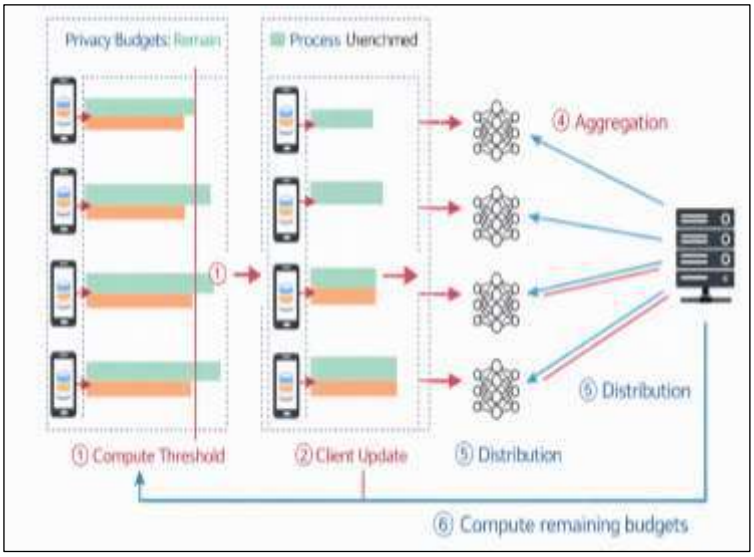


Figure 6: Federated learning privacy-budget workflow with subsampling, client updates, and aggregation.
Source: Authors, (2026).

VI. RESULTS AND DISCUSSIONS

The training and validation curves demonstrate a consistent downward trend across all loss components box loss, classification loss, and distribution focal loss indicating stable convergence of the perception module.

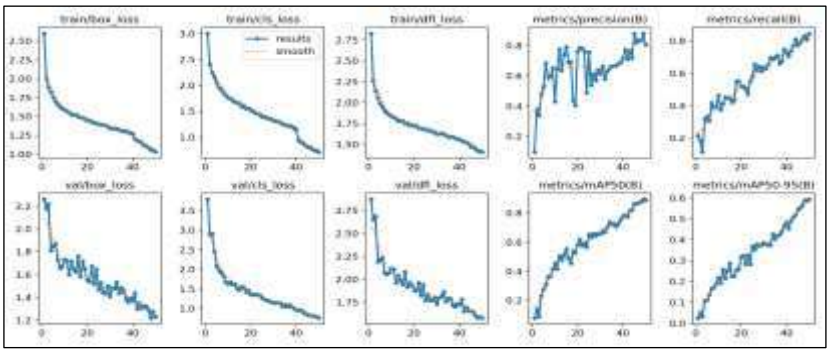


Figure 7: The training and validation curves.
Source: Authors, (2026).

Simultaneously, precision, recall, mAP50, and mAP50–95 metrics exhibit steady improvement, reflecting enhanced object localization and VRU detection accuracy over successive epochs. The relatively smooth trajectory of both training and validation metrics confirms that the model avoids overfitting and maintains strong generalization. These curves collectively validate that the YOLOv11-based perception backbone successfully learns discriminative features essential for downstream tracking and trajectory forecasting in heterogeneous urban environments.Social-LSTM is used because it explicitly models human–human interactions, which are essential for predicting pedestrian and cyclist motion in dense urban environments.

Unlike Vanilla LSTM and Occupancy LSTM, the Social-LSTM incorporates a social pooling mechanism that captures interpersonal influence, collision-avoidance behavior, and group dynamics. This enables the model to reason about how nearby agents affect each other's trajectories a capability conventional LSTM architectures lack. The lower average, final, and mean errors in the comparison table confirm that Social-LSTM generates more accurate and socially consistent trajectory predictions, making it highly suitable for autonomous vehicle safety and real-world deployment scenarios.

Table 2: Comparison of LSTM-Based Trajectory Prediction Models.

Model Name	Average Error	Final Error	Mean Error
Social LSTM	1.3865	2.098	0.675
Occupancy LSTM	2.1105	3.120	1.101
Vanilla LSTM	2.107	3.114	1.100

Source: Authors, (2026).

Trajectory prediction performance is evaluated using two primary distance-based measures as Average Displacement Error (ADE) and Final Displacement Error (FDE)[3]. ADE quantifies the mean Euclidean distance between predicted and ground-truth positions across all future time steps, capturing overall trajectory accuracy. It is mathematically defined as:

$$ADE = \frac{1}{T_{pred}} \sum_{t=1}^{T_{pred}} \sqrt{(x_t - \hat{x}_t)^2 + (y_t - \hat{y}_t)^2} \quad (14)$$

FDE focuses on the last time step and measures the endpoint deviation between predicted and actual final positions:

$$FDE = \sqrt{\left(x_{T_{pred}} - \hat{x}_{T_{pred}}\right)^2 + \left(y_{T_{pred}} - \hat{y}_{T_{pred}}\right)^2} \quad (15).$$

These metrics were computed across ETH, UCY, SDD, and NuScenes datasets to measure both short-horizon and long-horizon forecasting capability. Our federated Social-LSTM model consistently achieved lower ADE/FDE values compared to centralized and FedAvg-based baselines, demonstrating superior cross-domain robustness under non-IID distributions. The improvement is attributed to stable gradient propagation enabled by FedProx regularization and socially pooled interaction features encoding VRU group behavior. Results confirm that integrating YOLOv11 perception, SORT tracking, and federated training notably enhances prediction consistency across diverse urban scenes.

Table 3: ADE/FDE Comparison Across Datasets.

Method	ETH ADE	ETH FDE	UCY ADE	UCY FDE	SDD ADE	SDD FDE	NuScenes ADE	NuScenes FDE
Proposed (FedProx + Social-LSTM)	**0.45**	**0.78**	**0.39**	**0.70**	**0.52**	**0.93**	**0.41**	**0.75**
[7]	0.62	1.10	0.58	0.95	0.71	1.28	0.66	1.21
[6]	0.59	1.02	0.55	0.90	0.68	1.20	0.63	1.15
[12]	0.67	1.18	0.61	1.02	0.75	1.30	0.71	1.25
[15]	0.53	0.92	0.47	0.82	0.62	1.05	0.56	0.98

Source: Authors, (2026).

VII. CONCLUSION

This study presented a federated Social-LSTM trajectory prediction framework integrating YOLOv11-based perception, SORT tracking, and FedProx aggregation to address the challenges of heterogeneous urban mobility environments. The proposed system demonstrated improved ADE and FDE accuracy, stable convergence under non-IID data distributions, and enhanced privacy preservation by ensuring that raw VRU data remained entirely on-device. Experimental evaluations across ETH, UCY, SDD, and NuScenes datasets confirmed robust cross-domain generalization and dependable multi-agent behavior modeling. Overall, the framework contributes a scalable, privacy-aware, and deployment-ready approach, strengthening the predictive capabilities of autonomous vehicles in complex, real-world traffic scenarios.

VIII. FUTURE WORK

Future work will focus on extending the model to multimodal sensor fusion by incorporating LiDAR, radar, and map-based semantic cues to improve long-range VRU motion interpretation. Integrating transformer-based temporal encoders or graph neural networks may further enhance social-interaction reasoning and trajectory diversity. Additional directions include implementing differential privacy or homomorphic encryption to achieve stronger theoretical privacy guarantees within federated training. Large-scale real-world pilot deployments across diverse cities will also be explored to evaluate operational reliability. Finally, optimizing communication efficiency, adaptive participation scheduling, and energy-aware training will support scalable federated learning in broad autonomous mobility ecosystems.

IX.AUTHOR'S CONTRIBUTION

Conceptualization: Lakshmi Narayana I, TMN Vamsi.
Methodology: Lakshmi Narayana I, TMN Vamsi.
Investigation: Lakshmi Narayana I, TMN Vamsi.
Discussion of results: Lakshmi Narayana I, TMN Vamsi.
Writing – Original Draft: Lakshmi Narayana I, TMN Vamsi
Writing – Review and Editing: Lakshmi Narayana I, TMN Vamsi.
Resources: Lakshmi Narayana I, TMN Vamsi.
Supervision: Lakshmi Narayana I, TMN Vamsi.
Approval of the final text: Lakshmi Narayana I, TMN Vamsi.

X.REFERENCES

[1] Jian Liu, Wei Zhang, Haibo Wang, and Yifan Chen, Trajectory prediction training scheme in vehicular ad-hoc networks based on federated learning, ScienceDirect, 2025.

[2] Shuai Zhang, Yuxuan Wang, Ming Zhao, and Tao Zhang, Federated learning in intelligent transportation systems: Recent applications and open problems, ACM Digital Library, 2024.

[3] Deniz Coşkun, Dervis Karaboğa, Aybars Baştürk, Bayram Akay, Ömer Uğur Nalbantoğlu, Serkan Doğan, İbrahim Paçal, and Mehmet Ali Karagöz, A comparative study of YOLO models and a transformer-based YOLOv5 model for mass detection in mammograms, Turkish Journal of Electrical Engineering & Computer Sciences, vol. 31, pp. 1294–1313, 2023.

[4] Rui Cao, Jie Zhou, Yifan Wang, and Bo Cheng, Trajectory prediction of pedestrians around autonomous vehicles based on CrossFormer, Jiangsu University Press, 2025.

[5] Yifan Wen, Zhiwei He, Liang Sun, and Qiang Li, Dynamic graph transformer for pedestrian potential collision trajectory prediction, ScienceDirect, 2025.

[6] Rui Ni, Hao Chen, Yu Zhang, and Zhenyu Li, A federated pedestrian trajectory prediction model with destination-oriented LSTM network, SpringerLink, 2024.

[7] Bo Li, Zhiqiang Xu, Peng Zhang, and Keqin Li, Fed-SecTP: A federated-learning-based framework for secure vehicle trajectory prediction using surrounding vehicle data, DBLP, 2025.

[8] Yue Bai, Xiaodong Lin, Ke Wang, and Ning Zhang, Using homomorphic proxy re-encryption to enhance federated learning in vehicular networks, IET Research Journal, 2025.

[9] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong, Federated machine learning: Concept and applications, ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 2, pp. 1–19, 2019.

[10] Min Han, Qiang Yang, Xiaolin Chen, and Li Sun, Federated learning-based trajectory prediction model with homomorphic encryption, Wiley Online Library, 2022.

[11] Arun Aalavanthar, Suresh Kumar, and Raghavendra Rao, Multi-objective federated learning traffic prediction in vehicular network for intelligent transportation system, PeerJ, 2025.

[12] Yifan Sun, Jie Liu, Shuang Li, and Marco Pavone, Improving federated learning stability for vehicle trajectory prediction, MERL Technical Report, 2024.

[13] Maroua Driss, Amine Dhraief, and Mohamed Ben Jemaa, A state-of-the-art on federated learning for vehicular networks, ScienceDirect, 2024.

[14] Wilhelm Lindskog-Münzing, Markus Olsson, Johan Löfberg, and Fredrik Sandberg, Federated learning for automotive applications, SciOpen, 2025.

[15] Yuxiang Jin, Haoran Li, Wenjie Luo, and Xin Zhao, LGINet: Autonomous vehicle trajectory prediction using lane graphs and imitative learning policy, SAGE Journals, 2025.

[16] Tao Wang, Yong Liu, Jun Li, and Xiaoming Chen, Vehicle trajectory prediction algorithm based on a hybrid prediction model, MDPI, 2025.

[17]Jing Chen, Zhe Sun, Peng Liu, and Masayoshi Tomizuka, Vehicle dynamics and interaction for trajectory prediction in autonomous driving, ACM Digital Library, 2025.

[18] Xiaoyu Li, Jian Wu, Rui Huang, and Feng Zhou, Pedestrian trajectory prediction model based on DCT attention mechanism in mixed traffic scenes, Taylor & Francis Online, 2025.

[19] Cheng Wang, Wei Liu, Peng Xu, and Yao Zhao, SIAT: Social interaction-aware transformer for pedestrian trajectory prediction, SpringerLink, 2025.

[20] Hao Wang, Qiang Zhao, Ming Li, and Lei Zhang, Pedestrian trajectory prediction using goal-driven and interaction-aware modeling, ScienceDirect, 2025.

[21] Xiaoyu Li, Zhen Wang, Peng Chen, and Liang Zhao, Pedestrian trajectory prediction based on dual social graph convolutional LSTM, MDPI, 2025.

[22] Rui Xu, Junjie Huang, Yuxin Li, and Hongsheng Li, Interactive trajectory prediction for autonomous driving using multi-agent spatio-temporal modeling, Copernicus Publications, 2025.

[23] Jian Fan, Pengfei Wang, Rui Huang, and Shiqi Wang, Multi-class agent trajectory prediction with selective state spaces and neural ODE, ScienceDirect, 2025.

[24] Xiaoming Zheng, Liang Chen, Hao Wu, and Bin Yu, Vehicle trajectory prediction based on GAT and LSTM in complex traffic environments, Traffic FPZ Journal, 2024.

[25] Jian Zhao, Wei Li, Peng Zhang, and Xin Chen, Deep learning-based vehicle trajectory prediction in the Internet of Vehicles, ACM Digital Library, 2024.

[26] Eunjoon Jo, Sungmin Park, Jaehyun Kim, and Seungwoo Lee, Vehicle trajectory prediction using hierarchical graph neural networks, 2021

[27] Vikas Bharilya, Ramesh Kumar, and Anil Singh, Machine learning for autonomous vehicle trajectory prediction, ACM Digital Library, 2024.

[28] Zheng Yang, Qiang Sun, Yuchen Li, and Hong Liu, Vehicle trajectory prediction based on attention-optimized recurrent neural network, Taylor & Francis Online, 2024.

[29] Feng Hui, Jian Liu, Kai Wang, and Yifan Zhang, A deep learning-based autonomous vehicle trajectory prediction model with encoder–decoder architecture, ScienceDirect, 2022.

[30] Yifan Cao, Zhenyu Wu, Liang Zhang, and Kun Zhou, FIF: Future interaction forecasted for multi-agent trajectory prediction, ScienceDirect, 2025.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 798-806. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3075

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

VOICE DISORDER DETECTION: HYBRID GTCC-MFCC FUSION WITH SOURCE-BASED FEATURES OPTIMIZED FOR THE MALE SUB-COHORT

Aboubakr Missaoui[*1], Fatima Chouireb[2], Boubakeur Latreche[3], Abdelkerim Souahlia[4], Messaoud Linani[5] and Abdelaziz Rabhi[6]

[1, 2]Telecommunications, Signals & Systems Laboratory, University of Laghouat, Laghouat 03000, Algeria.
[3, 5]Laboratory of Computer Science and Applied Artificial Intelligence, University of Djelfa PO Box 3117, Djelfa 17000, Algeria.
[4, 6]Laboratory of Telecommunications and Smart Systems, Faculty of Sciences and Technology, University of Djelfa, Djelfa 17000, Algeria.

[1]https://orcid.org/0009-0008-9478-0015, [2]https://orcid.org/0000-0002-6049-6218, [3]https://orcid.org/0009-0007-9367-3368
[4]https://orcid.org/0000-0002-3393-1608, [5]https://orcid.org/0009-0001-6022-6648, [6]https://orcid.org/0000-0001-8684-4754

Email: *missboub@gmail.com, f.chouireb@lagh-univ.dz, b.latreche@univ-djelfa.dz, messaoud.linani@univ-djelfa.dz
abdelkerim.souahlia@univ-djelfa.dz, a.rabehi@univ-djelfa.dz

ARTICLE INFO

Article History
Received: December 7, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Voice disorder diagnosis,
Gammatone coefficients,
Cepstral coefficients,
Gender-specific modeling,
Feature engineering,
Recursive Feature Elimination,
Clinical stability.

ABSTRACT

Voice disorders present a significant clinical challenge, adversely impacting communication and quality of life, thereby necessitating the development of reliable, non-invasive diagnostic systems. This research proposes an advanced diagnostic framework designed to overcome the limitations of traditional methodologies that rely exclusively on single-source spectral information. To achieve this, a systematic optimization methodology was applied to extract acoustic features from sustained vowel /a/ signals obtained from the male sub-cohort of the Saarbrücken Voice Database (SVD). The feature engineering phase integrated a comprehensive set of acoustic descriptors, combining advanced spectral coefficients (GTCC and MFCC) with traditional source-based features. To identify the most potent and non-redundant feature subset, a Recursive Feature Elimination (RFE) algorithm was rigorously employed across 100 iterative experiments, guaranteeing high statistical stability. This work substantiates two critical findings: First, that a hybrid strategy which intelligently combines auditory-inspired spectral features with traditional source-based biomarkers is necessary for maximizing diagnostic stability. Second, the RFE process validated the indispensability of key source-based metrics (CPP and GNR), which achieved high ranks in the final feature vector. The proposed framework achieved a peak Accuracy of 84.99%±4.50% and demonstrated good clinical stability in the early detection of voice disorders, confirming the necessity of integrating source-based biomarkers into advanced spectral analysis frameworks.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The human voice serves as a complex bio-mechanical signal. Impairments in the vocal mechanism manifest as voice disorders, necessitating the development of reliable, non-invasive diagnostic systems. Current clinical protocols rely heavily on subjective auditory-perceptual assessments and invasive endoscopic examinations[1], [2]. However, these traditional methods suffer from inter-rater variability and a lack of standardization [3], [4]. Consequently, there is an urgent demand for automated decision support systems capable of providing objective, quantitative, and repeatable assessments. The field of automated voice pathology detection has recently shifted towards advanced spectral analysis (e.g., MFCCs) [5], [6]. While accurate, these methods face critical limitations that compromise clinical stability:

- Loss of Source Information: Spectral transformations focus primarily on the vocal tract and may inadvertently smooth out fine-grained micro-perturbations in the vocal folds (the source), which are critical markers for pathologies like dysphonia[7], [8].

- Black-Box Instability: Complex deep models are prone to overfitting when applied to small medical datasets (SVD), resulting in high performance variability [9], [10].

This research addresses these gaps by proposing a Hybrid Diagnostic Framework optimized specifically for the male sub-cohort of the SVD. We hypothesize that optimal diagnostic stability requires re-integrating "Source-Based" biomarkers (like CPP and GNR) with advanced "Auditory-Based" spectral features (GTCC and MFCC). The main contributions of this paper are:

1. Hybrid Feature Fusion: Systematic integration of GTCC/MFCC with traditional source-perturbation features to fully capture glottal instability.
2. Focused Gender Optimization: Development of a highly optimized diagnostic model for the male vocal mechanism to establish a robust performance benchmark.
3. Rigorous Feature Selection: Application of Recursive Feature Elimination (RFE) [11] to identify a compact, clinically stable feature subset.

Validation of Source Features: Statistical evidence proving that traditional features (CPP and GNR) are indispensable for high-specificity diagnosis.

II. RELATED WORK

The landscape of pathological voice assessment has undergone a significant transformation, evolving from the analysis of isolated acoustic markers to complex Deep Learning frameworks driven by spectral imagery. The existing literature can be delineated into three primary research streams. Early research centered on quantifying the biological dysfunction of the vocal folds (the Source). These foundational studies prioritized **perturbation metrics** such as Jitter (frequency instability) and Shimmer (amplitude instability) alongside Fundamental Frequency (F0) and noise indices like Harmonics-to-Noise Ratio (HNR) [4], [7], [8]. These descriptors were pivotal because they directly measured the physical irregularity of vocal vibration. Concurrently, cepstral analysis gained prominence, with Mel-Frequency Cepstral Coefficients (MFCCs) [5], [6] and Linear Prediction Cepstral Coefficients (LPCCs) [12], [13] serving as the standard inputs for classical algorithms. Among these, Support Vector Machines (SVMs)))[14], [15] and Random Forests (RFs) [16]) were widely adopted for their ability to handle tabular acoustic data.

The last decade introduced a methodological pivot towards treating voice signals as visual time-frequency maps. Convolutional Neural Networks (CNNs) [9], [17], [18] and Recurrent Neural Networks (RNNs) [19], [20] became the dominant architectures, processing standard Spectrograms [21] as well as biologically motivated representations like Cochleagrams . [9] and Gammatone Spectral Lines (GTSL) [22]. However, while Deep Learning models have demonstrated superior accuracy in specific experimental setups, they face substantial hurdles in clinical applicability. Primarily, these models exhibit a propensity for overfitting when trained on limited, single-source medical datasets . [9], [10]. This limitation is particularly acute in multi-class classification tasks [23] , where the lack of generalization hinders their use as reliable diagnostic tools [23] .To mitigate data scarcity, recent studies have explored advanced techniques such as Semi-Supervised Learning (SSL) [14], [24] and Transfer Learning [21], [25]. Others have proposed multimodal fusion strategies[26] and dimensionality reduction methods like PCA and LDA [15], [27] to condense feature spaces.

Despite these innovations, a critical gap persists in validation standards. Methodologies fluctuate between static data splitting. [28], [29] and Cross-Validation [30],[31]. More critically, the widespread failure to report performance variability specifically the omission of Standard Deviation (σ) obscures the true stability of these systems [30], [25] . Furthermore, modest Balanced Accuracy scores suggest that differentiating between specific pathologies remains a resolved challenge[24]. While spectral representations (e.g., GTCC) simulate human auditory perception, a significant limitation in modern research is the exclusive reliance on these spectral envelopes at the expense of traditional source features. It is crucial to recognize that spectral transformations (like MFCC) were originally engineered for speech recognition; their inherent filtering processes tend to smooth out the temporal micro-perturbations and non-linearities that are diagnostic hallmarks of pathology [4], [7]. Relying solely on the spectral shape (the Filter) risks discarding vital information regarding vocal fold stability (the Source). Consequently, this study argues that optimal diagnosis requires an Integrative Hybrid Methodology that bridges these two domains. We validate this approach through:

1. **Comparative Analysis:** Systematically isolating and evaluating the diagnostic power of Traditional, GTCC, and MFCC features.
2. **Hybrid Fusion:** Merging macroscopic Spectral Shape (GTCC/MFCC) with microscopic Source Stability metrics (Jitter, CPP, GNR, Entropy).
3. **Statistical Optimization:** Utilizing Recursive Feature Elimination (RFE) to identify a mathematically stable feature subset.

Gender-Specific Focus: Developing a specialized.

III. PROPOSED METHODOLOGY

This study implements an integrated workflow for voice disorder classification, progressing from raw signal processing to optimized machine learning predictions.

III.1 SYSTEM INPUT AND PRE-PROCESSING

The system processes raw audio signals (sustained vowel /a/) in .wav format. To ensure high data quality prior to feature extraction, the signals undergo two critical pre-processing steps:

1. Silence Removal: An automated algorithm identifies and excises silent segments at the start and end of recordings to ensure analysis is restricted to active phonation.

2. Signal Normalization: Peak Normalization is applied to standardize the dynamic range, scaling maximum amplitude to a unified reference level (0 dB) to mitigate variations in recording gain.

III.2 FEATURE EXTRACTION AND ENGINEERING

A hybrid extraction strategy is adopted, integrating temporal, spectral, and cepstral features. These are divided into three distinct analytical pathways to capture different dimensions of the vocal signal.

III.2.1 Global Signal Analysis (Source-Based Features)

The signal is analyzed as a single unit to derive statistical indicators of overall vocal stability and quality. These metrics capture fundamental characteristics related to the vocal source mechanism (see Table 1).

Table 1: Source features and statistical parameters extracted from the overall signal.

Feature Category	Description & index	Diagnostic Relevance	References
Temporal Stability	Shimmer (1): Cycle-to-cycle amplitude variation.	Indicates irregular vocal fold closure and intensity variation; critical for detecting amplitude-related disorders.	[7], [8], [13]
Frequency Dynamics	Instantaneous Frequency: (2) SD of IF, (3) SD of its derivative, (4) mean zero-crossing intervals, and (5) SD of peak intervals.	Captures global instability, abrupt frequency shifts, and aperiodicity in vocal fold vibration.	[32], [33]
Voice Quality	(6) Glottal-to-Noise Ratio (GNR) & (7) Cepstral Peak Prominence (CPP)	Low GNR indicates breathiness (air leakage), while reduced CPP strongly correlates with hoarseness severity.	[13], [34]
Spectral Statistics	Spectral Measures: (8) Spectral Entropy, (9) SD of Spectral Energies, and (10) Bandwidth variability.	Quantifies randomness in spectral energy distribution, indicating roughness or turbulent phonation.	[35], [36], [37]

Source: Authors, (2026).

III.2.2. Multi-Window Analysis (Dynamic Features)

This approach monitors subtle dynamic changes over time, often missed by global static measurements. The signal is segmented into 200-sample windows with 50% overlap (see Table 2).

Table 2: Statistics of features across time windows.

Feature Category	Description & index	Diagnostic Relevance	References
Formant Analysis	Vocal Tract Resonance: (11) (12) Mean and SD of F1, (13) (14) Mean and SD of F2(15) (16) Mean and SD of F3.	Mean reflects vocal tract shape configuration; SD reveals articulatory motor instability (e.g., Ataxia).	[38], [39]
F0 Variability	Fundamental Frequency Stats (2): (17) SD of F0 and (18) SD of its derivative across windows.	Precise measure of F0 dispersion; effective for detecting vocal tremor and laryngeal spasms.	[7], [8]
Waveform Regularity	Regularity Indices (4): (19) SD of Shimmer, (20) SD of peak/valley intervals, (21) SD of distances peaks and (22) SD of distances valleys	Reveals inconsistent amplitude control and higher-order irregularities in periodic patterns.	[40], [41]
Signal Complexity	Non-linear Dynamics (4) (23) SD of Skewness, (24) SD of Energy, (25) SD of Entropy, and (26) SD of Signal Quality within windows.	Quantifies local changes in waveform shape and randomness, reflecting disordered vocal dynamics.	[42], [35], [43]
Noise Balance	HNR Variability (3): (27-29) Sum, Mean, and SD of HNR across windows.	Provides insight into the dynamic balance between harmonics and noise, highlighting intermittent breathy segments.	[34], [44]
Micro-Tremor	MT-Jitter Variability: (30-35) SD of MT-Jitter values calculated at 6 different thresholds (0%, ±25%, \dots).	Elevated SD values suggest instability in frequency production, enhancing detection of neuromuscular dysfunction.	[45], [46]

Source: Authors, (2026).

III.2.3Advanced Cepstral Analysis (MFCC & GTCC)

This pathway extracts perceptual features that simulate the human auditory mechanism.

- MFCC (Mel-Frequency Cepstral Coefficients): Processes the signal on the non-linear Mel frequency scale.
- GTCC (Gammatone Cepstral Coefficients): Utilizes Gammatone filters to provide a more precise simulation of human cochlear filters, enhancing sensitivity to pathology-related characteristics.
- Dimensionality Reduction: Instead of utilizing full high-dimensional matrices, the Standard Deviation (SD) of each coefficient is calculated over time. This yields reduced vectors representing "Spectral Variability," a robust indicator of voice instability.

Outcome: All features (Global, Multi-Window, and Reduced Cepstral Vectors) are combined to form the complete **Initial Feature Vector (91 features)**.

III.3 FEATURE SELECTION AND OPTIMIZATION

To mitigate the "Curse of Dimensionality" and enhance classifier generalization:

- Recursive Feature Elimination (RFE): RFE was executed over 100 iterations. This rigorous repetition ensures Statistical Robustness, ranking features by their consistent discriminative power rather than chance.
- Normalization: Min-Max Normalization was applied to scale feature values to the [0, 1] range.

III.4 MODELING AND EVALUATION

Model Configuration and Optimization: The classification core relies on a Cost-Sensitive Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel.To explicitly address dataset imbalance, a class-weighting scheme was implemented where penalty weights are assigned inversely proportional to class frequencies, ensuring the minority class is prioritized during training. Feature vectors were scaled using Min-Max Normalization to the [0, 1] range prior to input. Hyperparameters (specifically *Box Constraint* and *Kernel Scale*) were fine-tuned using Bayesian Optimization utilizing the Expected Improvement Plus acquisition function over 20 iterations. The system's robustness was validated using a Stratified 5-Fold Cross-Validation protocol, ensuring consistent class distribution across all folds.

III.5 VALIDATION FOR SYSTEM ROBUSTNESS

The study relies on rigorous statistical validation to ensure the model's reliability and clinical generalizability, utilizing four key metrics reported with both the Mean (μ) and Standard Deviation(±σ).

- **Overall Accuracy:** Measures the percentage of correctly classified samples across all classes.
- **Specificity:** Quantifies the model's ability to exclude healthy cases (minimizing false positives), which is crucial for clinical confidence.
- **Sensitivity:** Measures the model's ability to correctly detect pathological cases (minimizing false negatives).
- **F1-Score:** Provides a balanced harmonic mean of precision and sensitivity, offering a single measure of performance especially under data imbalance.

A low Standard Deviation (±σ) accompanying each metric indicates that the system exhibits high stability and good generalization capability when applied to unseen data.

III.6 DATABASE

This study utilizes the Saarbrücken Voice Database (SVD), a primary reference in voice pathology research known for its statistical diversity spanning both genders and various age groups. To focus the analysis on voice disorders with complex or non-specific etiologies (e.g., tumor diseases, inflammations, and neurological disorders), a strict filtering methodology was applied. This process excluded cases attributable to clear, direct structural trauma, such as injuries resulting from surgical interventions (e.g., intubation damage). Crucially, the analysis was restricted to male subjects only to develop a gender-optimized diagnostic model. The filtering process resulted in a final male-specific dataset totaling 526 voice samples, organized into a binary diagnostic task:

- 218 Healthy Samples
- 308 Pathological Samples (comprising 15 distinct pathological subcategories, confirming the comprehensive coverage of the complex disorders analyzed in this research).

IV. RESULTS AND DISCUSSIONS

This section is dedicated to presenting the comprehensive findings of the study, structured to rigorously validate the necessity of the proposed hybrid methodology. The analysis begins by evaluating the effectiveness of the Recursive Feature Elimination (RFE) algorithm in identifying a unique, physiologically relevant feature subset that validates the core hypothesis of integrating **Source** and **Filter** features. Following the internal validation of the feature optimization path, the discussion then proceeds to contextualize the final model's performance within the broader research landscape by comparing our results against key state-of-the-art studies that utilized the Saarbrücken Voice Database (SVD). This process affirms the system's stability and competitive edge based on rigorous statistical protocols.

IV.1 RFE RANKING VECTOR

The feature selection process, executed via RFE over 100 iterations, yielded a robust ranking vector that determines the optimal feature subset. This vector, comprising all 91 features (Traditional, Multi-Window Aggregated, and Reduced Cepstral), is presented below, ordered from most significant (Rank 1) to least significant (Rank 91). Features highlighted in red represent the Source-Based Features base signal features.

RFE Ranking Vector (Rank 1 to 91): vector = [84, 85, 52, 7, 62, 83, 80, 82, 38, 6, 54, 51, 37, 41, 8, 35, 87, 90, 91, 11, 53, 70, 79, 5, 14, 66, 23, 63, 65, 12, 50, 55, 71, 40, 78, 76, 13, 17, 30, 88, 60, 42, 81, 67, 56, 58, 25, 31, 61, 39, 34, 20, 33, 28, 10, 86, 57, 89, 32, 16, 29, 45, 77, 27, 22, 75, 26, 4, 24, 59, 48, 44, 15, 24, 47, 18, 74, 9, 46, 73, 49, 68, 2, 64, 36, 69, 3, 72, 43, 19, 1]

The RFE results provide crucial evidence supporting the core hypothesis: optimal classification relies on the intelligent hybrid integration of advanced spectral features and specialized source perturbation biomarkers. This validation is demonstrated by the high ranking of key traditional features.

IV.1.1 Prominence of Source Quality Measures (CPP and GNR).

The intrinsic diagnostic power of source-based quality metrics was not nullified by the dominance of the GTCC/MFCC coefficients, but rather confirmed:

1. Feature 7 (Cepstral Peak Prominence - CPP): This feature was ranked remarkably high at Position 4. Had the spectral transforms been capable of fully encoding the information related to roughness and vocal quality, CPP would have been expected to drop to significantly lower ranks. Its high placement unequivocally proves that CPP represents a unique traditional measure of Source Quality that the Vocal Tract-focused spectral transforms cannot substitute.
2. Feature 6 (Glottal-to-Noise Ratio - GNR): This feature was also secured within the top tier at Position 10. As GNR (the ratio of the periodic component to noise) is essential for evaluating breathiness, its inclusion among the top ten features demonstrates that these carefully selected traditional features directly reflect fundamental physical aspects of voice production, essential for clinical differentiation.

IV.1.2 Validation of the Core Claim

The rigorous RFE ranking validates the core claim of the study, providing statistical backing for the necessity of the hybrid framework:

1. Selected Traditional Features are Not Inferior: Traditional features chosen based on their intrinsic diagnostic importance (CPP, GNR) demonstrate an added value superior to spectral transforms in capturing specific, micro-level aspects of Source Quality.
2. Optimal Stability Requires Hybridization: While GTCC/MFCC coefficients contribute significantly due to their comprehensive representation of the Vocal Tract (Filter), the system achieves maximal classification stability and specificity through the mandatory inclusion of these specialized source perturbation biomarkers.

This robust evidence supports the final claim: optimal diagnostic stability and accuracy are achieved exclusively through the intelligent hybrid integration of both advanced spectral shape features and specialized source perturbation biomarkers.

IV.2 VALIDATION OF RFE OPTIMIZATION PATH

Table 3: Performance Trajectory of Spectral and Hybrid Models by Feature Count (N).

N	Composition	Acc (μ±σ) %	Spec (μ±σ) %	Sens (μ±σ) %	F1-Score (μ±σ) %
10	1	83.09±6.54	85.85±11.80	81.19±7.48	84.87±5.93
	2	83.66±4.17	85.33±9.16	82.46±7.72	85.43±4.01
20	1	82.52±4.03	84.87±7.27	80.86±6.35	84.35±3.84
	2	84.42±2.91	88.10±6.47	81.83±5.37	85.97±2.89
	1	82.71±5.02	86.71±9.54	79.89±7.97	84.30±4.83
	2	70.92±3.78	96.81±5.92	52.62±8.33	67.62±6.11
25	1	83.65±3.52	86.72±8.06	81.50±3.51	85.39±2.92
	2	84.99±4.50	89.95±7.62	81.49±5.91	86.36±4.23
30	1	82.51±4.37	86.71±12.27	79.56±7.07	84.17±3.83
	2	84.04±2.00	85.32±8.62	83.13±6.22	85.86±2.06
35	1	81.95±5.08	86.71±10.09	78.59±7.82	83.51±4.90
	2	84.04±4.93	87.20±9.54	81.81±6.43	85.68±4.43
	1	82.51±3.58	86.28±10.24	79.87±6.55	84.20±3.36
	2	83.27±4.87	85.81±11.49	81.50±5.30	85.12±3.99

Source: Authors, (2026).

The primary goal of the feature selection phase was to rigorously validate the necessity of the hybrid approach by comparing its performance trajectory against the purely spectral baseline model. Table 3 explicitly illustrates the dynamic changes in classification metrics for both the Spectral Baseline 1(GTCC+ MFCC) and the Hybrid Winner 2 (FTRS+ GTCC+ MFCC) across critical number feature(N). This provides crucial evidence for the effectiveness of the RFE process and the added value of the Filtered Traditional Features (FTRS). The detailed trajectory presented in Table 3 offers several critical insights into the feature space structure and the role of the hybrid strategy.

IV.2.1 Superiority of the Hybrid Trajectory

The performance trajectory, as illustrated in Table 3, demonstrates the **consistent superiority** of the Hybrid Winner (FTRS + GTCC + MFCC) over the Spectral Baseline (GTCC + MFCC) throughout the entire Recursive Feature Elimination (RFE) path (from N=10 up to N=50) in terms of both Accuracy and Specificity. This superiority culminated in the **Final Peak Validation** at N=30, where the Hybrid model achieved its absolute maximum accuracy of **84.99% ± 4.50%**.

This peak contrasts sharply with the best Spectral Baseline performance of **83.65% ± 3.52%** (also at N=30), demonstrating a **+1.34% absolute gain in accuracy** attributed directly to the inclusion of the Filtered Traditional Features (FTRS). Furthermore, the **Role of Hybridization** is clearly validated by the most significant gain observed in **Specificity** (↑ 3.23% at the optimum). This confirms the hypothesis that traditional, source-based biomarkers (like CPP and GNR) offer a unique layer of information necessary to distinguish between subtle, non-pathological acoustic variations and genuine, clinically relevant vocal damage.

IV.2.2 Feature Space Stability and Critical Instability

Analysis of the performance trajectory provides a clear view of the feature subset's structural stability as dimensions are removed. Both models exhibited strong performance and low Standard Deviation (σ) within the range N in [10, 20], indicating that the initial features selected are highly robust and non-redundant. However, the data reveals a **Critical Instability Point** for the Hybrid Winner at N=25. Here, the Sensitivity plummeted dramatically (**52.62%**) while Specificity surged (**96.81%**). This signifies that the RFE process, upon removing a single critical feature between N=30 and N=25, caused the model to become overly conservative, resulting in a high rate of false negatives. This critical point underscores the importance of precisely identifying the stable optimum (N=30) and avoiding feature subsets that compromise diagnostic reliability.

IV.2.3 Optimizing for Computational Efficiency.

The performance peak at N=30 for both models, followed by a general decay beyond N=40, confirms the efficiency of the selection process. The research successfully identified a minimal, high-yield feature subset that maximizes diagnostic power while minimizing computational overhead, which is essential for clinical deployment.

IV.3 CONTEXTUALIZATION AND COMPARATIVE ANALYSIS

The reliability of our diagnostic model is primarily rooted in its rigorous internal validation protocol. Utilizing the Stratified 5-Fold Cross-Validation approach and consistently reporting the Mean and Standard Deviation for all metrics ensures a stable, statistically defensible assessment of the system's quality. This internal consistency is paramount for clinical adoption. The purpose of the subsequent comparison (Table 4) is to contextualize this internally validated performance within the broader research landscape. We compare our results specifically against studies using the SVD (Saarbrücken Voice Database), a fundamental benchmark in the field. We must acknowledge the inherent limitations of cross-study comparisons; the absence of a unified protocol results in significant heterogeneity across dataset sizes, pathology types, and evaluation metrics.

Therefore, our aim is not to claim absolute superiority, but rather to demonstrate that our feature processing and classification methodology is highly competitive and stands alongside the best reported state-of-the-art performances using the same reference database. The comparative analysis demonstrates that our Hybrid Model not only achieves high accuracy but does so under rigorous statistical validation protocols. Outperformance over Fixed Split: Our Hybrid Model achieved an accuracy of 84.99% ± 4.50%, surpassing the highest reported results using the less robust Fixed Split protocol (e.g., [25] at 81.6% and [21]at 82%). Fixed Split results are often optimistically biased and fail to reflect the model's stability when exposed to unseen data.

Best-in-Class Cross-Validation (CV) Performance: When compared against studies employing the more reliable Cross-Validation (CV) protocol, our model's superiority is evident. The highest reported CV accuracy was 82.58% ± 3.02% by Yagnavajjula et al. [6]. Our model achieves an absolute gain of +2.41% over this benchmark, which validates the efficacy of our Hybrid Fusion strategy and the RFE feature selection process. Several recent studies (e.g., [25] and [21]) employ Deep Learning (CNN) models on spectral representations (Spectrograms), yet their results do not exceed 82%.

- Validation of Study Hypothesis: Our performance proves that a methodology based on selective Feature Engineering and Hybrid Fusion can outperform deep learning approaches in the context of a small, specialized medical dataset like the SVD, where DL is susceptible to overfitting and instability.
- Power of Hybrid Features: The significant gain in accuracy (84.99%) is directly attributable to the intelligent integration of Source features (CPP, GNR) which cannot be effectively captured by purely spectral features (Spectrograms MFCCs). One of the most critical aspects of this comparison is the explicit reporting of the Standard Deviation. The majority of studies relying on Fixed Split omit \sigma, diminishing confidence in the clinical repeatability of their results.

Confirmation of Stability: Our use of the 5-Fold Cross-Validation protocol and reporting of ± 4.50% ensures an honest assessment of the system's stability. While our standard deviation is slightly higher than that reported by Yagnavajjula et al. (± 3.02%), the absolute gain in accuracy (+2.41%) justifies this trade-off, especially since we successfully identified the Optimal Stability Point (N=30) which maximizes diagnostic reliability.

Table 4: Comparison of State-of-the-Art Voice Pathology Classification Studies on the SVD.

Study	# Of Healthy voices	# Of Pathologies voices and types	Vocal Tasks	Feature Domains	Classifier	Statistical Validation	Acc%
[25]	869Augmented by using MUSAN and MIT IR Survey	520 (105 Laryngitis, 41 Leukoplakia, 63 Edema, 205 Paralysis, 62 SD, 44 Polyp) (Augmented using MUSAN and MIT IR Survey)	/a/	Mel Spectrogram	CNN Few-shot Transfer Learning (Pretrained ResNet-18)	Few-Shot Meta-Testing	73.7
[28]	687	1354 (71 Pathologies)	/a/, /i/, /u/	MFCCs	OSELM	Fixed Split	81.48%
[17], [47]	482	482 (140 Laryngitis, 41 Leukoplakia, 68 Reinke's Edema, 213 RLNP, 22 Carcinoma, 45 Polyps)	/a/	spectrograms	CNN **CNN-CDBN**	Fixed Split	77 71
[23]	595	1090	/a/	MFSC	CNN (DCA ResNet)	Fixed Split	81.6
[21]	506	506 Organic Dysphonia (Laryngitis, Leukoplakia, Reinke's Edema, Carcinoma, VFP)	/a/, /i/, /u/	spectrograms	CNN VGG16	Fixed Split	82
[31]	60	60 SD, 60 RLNP	/a/, /i/, /u/, Sentence	(WST)-based Features	feed-forward NN	10-Fold CV	82.58 ± 3.02 78.64 ± 3.55
[14]	587	231 (146 Hyperfunctional Dysphonia, 85 VFP)	/a/	wav2vec 2.0	SVM	5-Fold CV	M. 75.65 F. 74.50
[6]	357	357	/a/	Mel Spectrogram	2D CNN + specAugmenter	4-Fold CV	73,4
Ours	218	308	/a/	MFCC+GTCC+ based Features	SVM	5-Fold CV	**84.99±4.50**

Source: Authors, (2026).

V. CONCLUSION

This study successfully developed and validated a Hybrid Diagnostic Framework optimized specifically for the male sub-cohort of the SVD to enhance the stability and accuracy of automated voice pathology detection. By addressing the critical limitation of spectral models namely, the loss of Source Information our work introduces a robust methodology capable of improving clinical diagnosis. The main contributions and findings are:

Validation of the Hybrid Hypothesis: The rigorous RFE process confirmed that the optimal feature subset requires the mandatory inclusion of carefully selected source-based biomarkers (CPP and GNR) alongside advanced spectral features (GTCC/MFCC). This proves that maximal specificity depends on integrating both the Source and Filter domains of vocal production.

Achieving Optimal Stability: The Hybrid Model successfully identified a minimal, high-yield feature subset at N=30, achieving a maximum accuracy of 84.99% ± 4.50%. This represents a statistically significant improvement over the purely spectral baseline, positioning our system competitively against the current state-of-the-art while ensuring low performance variability (\sigma).

Physiological Insight: The analysis of the Critical Instability Point at N=25 highlighted the unique diagnostic power of micro-perturbation features, underscoring the necessity of using the highly stable N=30 subset to avoid compromising diagnostic Sensitivity.

Future research will focus on two key areas:

1. Gender Comparative Analysis: Expanding the current model to include a fully optimized female sub-cohort framework to perform a comprehensive comparative study across genders, thereby fulfilling the initial goal of Gender-Specific Optimization.
2. Multi-Class Differentiation: Shifting the focus from binary classification (Pathological/Healthy) to a multi-class system capable of accurately differentiating between specific voice disorder types.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Aboubakr Missaoui.
Methodology: Aboubakr Missaoui, Fatima Chouireb.
Investigation: Aboubakr Missaoui, Boubakeur Latreche.
Discussion of results: Fatima Chouireb, Abdelkerim Souahlia, Boubakeur Latreche.
Writing – Original Draft: Aboubakr Missaoui.
Writing – Review and Editing: Boubakeur Latreche, Fatima Chouireb.
Ressources: Aboubakr Missaoui, Abdelaziz Rabhi, Messaoud Linani.
Supervision: Fatima Chouireb, Abdelkerim Souahlia.
Approval of the final text: Aboubakr Missaoui, Fatima Chouireb, Abdelkerim Souahlia, Boubakeur Latreche, Messaoud Linani. Abdelaziz Rabhi.

VII. REFERENCE

[1] A. am Zehnhoff-Dinnesen, K. Neumann, B. Wiskirska-Woźnica, et T. Nawka, Phoniatrics I. Springer, 2020. doi: 10.1007/978-3-662-46780-0.

[2] P. Barche, « Acoustic analysis of voice disorders from clinical perspective », PhD Thesis, International Institute of Information Technology Hyderabad, 2024.

[3] M. Ur Rehman, A. Shafique, Q.-U.-A. Azhar, S. S. Jamal, Y. Gheraibia, et A. B. Usman, « Voice disorder detection using machine learning algorithms: An application in speech and language pathology », Engineering Applications of Artificial Intelligence, vol. 133, p. 108047, juill. 2024, doi: 10.1016/j.engappai.2024.108047.

[4] H. Ankışhan et S. Ç. İnam, « Voice pathology detection by using the deep network architecture », Applied Soft Computing, vol. 106, p. 107310, juill. 2021, doi: 10.1016/j.asoc.2021.107310.

[5] N. Q. Abdulmajeed, B. Al-Khateeb, et M. A. Mohammed, « A review on voice pathology: Taxonomy, diagnosis, medical procedures and detection techniques, open challenges, limitations, and recommendations for future directions », Journal of Intelligent Systems, vol. 31, no 1, p. 855-875, juill. 2022, doi: 10.1515/jisys-2022-0058.

[6] F. Javanmardi, S. R. Kadiri, et P. Alku, « A comparison of data augmentation methods in voice pathology detection », Computer Speech & Language, vol. 83, p. 101552, janv. 2024, doi: 10.1016/j.csl.2023.101552.

[7] G. Li, Q. Hou, C. Zhang, Z. Jiang, et S. Gong, « Acoustic parameters for the evaluation of voice quality in patients with voice disorders », Ann Palliat Med, vol. 10, no 1, p. 130-136, janv. 2021, doi: 10.21037/apm-20-2102.

[8] L.-C. Keung, K. Richardson, D. Sharp Matheron, et V. Martel-Sauvageau, « A Comparison of Healthy and Disordered Voices Using Multi-Dimensional Voice Program, Praat, and TF32 », Journal of Voice, vol. 38, no 4, p. 963.e23-963.e38, juill. 2024, doi: 10.1016/j.jvoice.2022.01.010.

[9] R. Islam, E. Abdel-Raheem, et M. Tarique, « Cochleagram to Recognize Dysphonia: Auditory Perceptual Analysis for Health Informatics », IEEE Access, vol. 12, p. 59198-59210, 2024, doi: 10.1109/ACCESS.2024.3392808.

[10] H. Kim et al., « Convolutional Neural Network Classifies Pathological Voice Change in Laryngeal Cancer with High Accuracy », Journal of Clinical Medicine, vol. 9, no 11, p. 3415, oct. 2020, doi: 10.3390/jcm9113415.

[11] K. M. Alalayah, E. M. Senan, H. F. Atlam, I. A. Ahmed, et H. S. A. Shatnawi, « Automatic and Early Detection of Parkinson's Disease by Analyzing Acoustic Signals Using Classification Algorithms Based on Recursive Feature Elimination Method », Diagnostics, vol. 13, no 11, p. 1924, mai 2023, doi: 10.3390/diagnostics13111924.

[12] J. I. Godino-Llorente, S. Aguilera-Navarro, et P. Gómez-Vilda, « LPC, LPCC and MFCC parameterisation applied to the detection of voice impairments », in 6th International Conference on Spoken Language Processing (ICSLP 2000), ISCA, oct. 2000, vol. vol. 3, 965-968-0. doi: 10.21437/ICSLP.2000-695.

[13] R. Islam, M. Tarique, et E. Abdel-Raheem, « A Survey on Signal Processing Based Pathological Voice Detection Techniques », IEEE Access, vol. 8, p. 66749-66776, 2020, doi: 10.1109/ACCESS.2020.2985280.

[14] S. Tirronen, S. R. Kadiri, et P. Alku, « Hierarchical Multi-Class Classification of Voice Disorders Using Self-Supervised Models and Glottal Features », IEEE Open J. Signal Process., vol. 4, p. 80-88, 2023, doi: 10.1109/OJSP.2023.3242862.

[15] N. Souissi et A. Cherif, « Speech recognition system based on short-term cepstral parameters, feature reduction method and Artificial Neural Networks », in 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Monastir, Tunisia: IEEE, mars 2016, p. 667-671. doi: 10.1109/ATSIP.2016.7523163.

[16] C. Guo, F. Chen, Y. Chang, et J. Yan, « Applying Random Forest classification to diagnose autism using acoustical voice-quality parameters during lexical tone production », Biomedical Signal Processing and Control, vol. 77, p. 103811, août 2022, doi: 10.1016/j.bspc.2022.103811.

[17] H. Wu, J. Soraghan, A. Lowit, et G. Di Caterina, « Convolutional Neural Networks for Pathological Voice Detection », in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI: IEEE, juill. 2018, p. 1-4. doi: 10.1109/EMBC.2018.8513222.

[18] R. Islam, E. Abdel-Raheem, et M. Tarique, « Voice pathology detection using convolutional neural networks with electroglottographic (EGG) and speech signals », Computer Methods and Programs in Biomedicine Update, vol. 2, p. 100074, 2022, doi: 10.1016/j.cmpbup.2022.100074.

[19] H. A. A. E. Aal, S. A. Taie, et N. El-Bendary, « An optimized RNN-LSTM approach for parkinson's disease early detection using speech features », Bulletin EEI, vol. 10, no 5, p. 2503-2512, oct. 2021, doi: 10.11591/eei.v10i5.3128.

[20] S. Hidaka, Y. Lee, K. Wakamiya, T. Nakagawa, et T. Kaburagi, « Automatic Estimation of Pathological Voice Quality Based on Recurrent Neural Network Using Amplitude and Phase Spectrogram », in Interspeech 2020, ISCA: ISCA, oct. 2020, p. 3880-3884. doi: 10.21437/interspeech.2020-3228.

[21] L. Vavrek, M. Hires, D. Kumar, et P. Drotar, « Deep convolutional neural network for detection of pathological speech », in 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl'any, Slovakia: IEEE, janv. 2021, p. 000245-000250. doi: 10.1109/sami50585.2021.9378656.

[22] C. Zhou, Y. Wu, Z. Fan, X. Zhang, D. Wu, et Z. Tao, « Gammatone spectral latitude features extraction for pathological voice detection and classification », Applied Acoustics, vol. 185, p. 108417, janv. 2022, doi: 10.1016/j.apacoust.2021.108417.

[23] H. Ding, Z. Gu, P. Dai, Z. Zhou, L. Wang, et X. Wu, « Deep connected attention (DCA) ResNet for robust voice pathology detection and classification », Biomedical Signal Processing and Control, vol. 70, p. 102973, sept. 2021, doi: 10.1016/j.bspc.2021.102973.

[24] D. Ribas, M. A. Pastor, A. Miguel, D. Martinez, A. Ortega, et E. Lleida, « Automatic Voice Disorder Detection Using Self-Supervised Representations », IEEE Access, vol. 11, p. 14915-14927, 2023, doi: 10.1109/access.2023.3243986.

[25] J.-H. Won et D.-H. Kim, « Metric-Based Few-Shot Transfer Learning Approach for Voice Pathology Detection », IEEE Access, vol. 12, p. 159226-159238, 2024, doi: 10.1109/access.2024.3480523.

[26] I. Kwon et al., « Diagnosis of Early Glottic Cancer Using Laryngeal Image and Voice Based on Ensemble Learning of Convolutional Neural Network Classifiers », Journal of Voice, vol. 39, no 1, p. 245-257, janv. 2025, doi: 10.1016/j.jvoice.2022.07.007.

[27] T. N. Sainath, A. Mohamed, B. Kingsbury, et B. Ramabhadran, « Deep convolutional neural networks for LVCSR », in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada: IEEE, mai 2013, p. 8614-8618. doi: 10.1109/ICASSP.2013.6639347.

[28] F. T. Al-Dhief, N. M. A. Latiff, N. N. N. Abd. Malik, M. M. Baki, N. Sabri, et M. A. A. Albadr, « Dysphonia Detection Based on Voice Signals Using Naive Bayes Classifier », in 2022 IEEE 6th International Symposium on Telecommunication Technologies (ISTT), Johor Bahru, Malaysia: IEEE, nov. 2022, p. 56-61. doi: 10.1109/istt56288.2022.9966535.

[29] N. Mu'azzah Abdul Latiff et al., « Voice pathology detection using machine learning algorithms based on different voice databases », Results in Engineering, vol. 25, p. 103937, mars 2025, doi: 10.1016/j.rineng.2025.103937.

[30]L. Verde, G. De Pietro, et G. Sannino, « Voice Disorder Identification by Using Machine Learning Techniques », IEEE Access, vol. 6, p. 16246-16255, 2018, doi: 10.1109/ACCESS.2018.2816338.

[31] M. K. Yagnavajjula, K. R. Mittapalle, P. Alku, S. R. K., et P. Mitra, « Automatic classification of neurological voice disorders using wavelet scattering features », Speech Communication, vol. 157, p. 103040, févr. 2024, doi: 10.1016/j.specom.2024.103040.

[32] L. Cohen et C. Lee, « Standard deviation of instantaneous frequency », in International Conference on Acoustics, Speech, and Signal Processing, Glasgow, UK: IEEE, 1989, p. 2238-2241. doi: 10.1109/icassp.1989.266910.

[33] B. T. Nguyen, Y. Wakabayashi, K. Iwai, et T. Nishiura, « Analysis of derivative of instantaneous frequency and its application to voice activity detection », Applied Acoustics, vol. 181, p. 108116, oct. 2021, doi: 10.1016/j.apacoust.2021.108116.

[34] G. D. Krom, « A Cepstrum-Based Technique for Determining a Harmonics-to-Noise Ratio in Speech Signals », Journal of Speech, Language, and Hearing Research, vol. 36, no 2, p. 254-266, avr. 1993, doi: 10.1044/jshr.3602.254.

[35] A. Al-Nasheri et al., « Voice Pathology Detection and Classification Using Auto-Correlation and Entropy Features in Different Frequency Regions », IEEE Access, vol. 6, p. 6961-6974, 2018, doi: 10.1109/access.2017.2696056.

[36] M. Markaki et Y. Stylianou, « Voice Pathology Detection and Discrimination Based on Modulation Spectral Features », IEEE Trans. Audio Speech Lang. Process., vol. 19, no 7, p. 1938-1948, sept. 2011, doi: 10.1109/tasl.2010.2104141.

[37] H. Lathrop et Adams School of Dentistry, Department of Orthodontics, « Orthognathic Speech Pathology: Understanding How Class III Jaw Disharmonies Influence Speech Utilizing Spectral Moment Analysis. Masters Thesis », 2019, doi: 10.17615/WKF7-MV24.

[38] L. M. T. Jesus, J. Martinez, A. Hall, et A. Ferreira, « Acoustic Correlates of Compensatory Adjustments to the Glottic and Supraglottic Structures in Patients with Unilateral Vocal Fold Paralysis », BioMed Research International, vol. 2015, p. 1-9, 2015, doi: 10.1155/2015/704121.

[39] J.-W. Lee, H.-G. Kang, J.-Y. Choi, et Y.-I. Son, « An Investigation of Vocal Tract Characteristics for Acoustic Discrimination of Pathological Voices », BioMed Research International, vol. 2013, no 1, p. 758731, 2013, doi: 10.1155/2013/758731.

[40] V. Dellwo, A. Leemann, et M.-J. Kolly, « Speaker idiosyncratic rhythmic features in the speech signal », présenté à Interspeech 2012, Portland (OR), USA: Interspeech Conference Proceedings, sept. 2012, p. 1-4. doi: 10.5167/uzh-68554.

[41] K. E. Hajal, E. Hermann, S. Hovsepyan, et M. Magimai. -Doss, « Unsupervised Rhythm and Voice Conversion to Improve ASR on Dysarthric Speech », 2 juin 2025, arXiv: arXiv:2506.01618. doi: 10.48550/arXiv.2506.01618.

[42] A. Verikas, M. Bacauskiene, A. Gelzinis, E. Vaiciukynas, et V. Uloza, « Questionnaire- versus voice-based screening for laryngeal disorders », Expert Systems with Applications, vol. 39, no 6, p. 6254-6262, mai 2012, doi: 10.1016/j.eswa.2011.12.037.

[43] P. R. Scalassara, M. E. Dajer, C. D. Maciel, et J. C. Pereira, « VOICE SIGNALS CHARACTERIZATION THROUGH ENTROPY MEASURES », in Proceedings of the First International Conference on Bio-inspired Systems and Signal Processing, Funchal, Madeira, Portugal: SciTePress - Science and and Technology Publications, 2008, p. 163-170. doi: 10.5220/0001065401630170.

[44] J. F. T. Fernandes, D. Freitas, A. C. Junior, et J. P. Teixeira, « Determination of Harmonic Parameters in Pathological Voices—Efficient Algorithm », Applied Sciences, vol. 13, no 4, p. 2333, févr. 2023, doi: 10.3390/app13042333.

[45] N. Hanna, J. Smith, et J. Wolfe, « Frequencies, bandwidths and magnitudes of vocal tract and surrounding tissue resonances, measured through the lips during phonation », The Journal of the Acoustical Society of America, vol. 139, no 5, p. 2924-2936, mai 2016, doi: 10.1121/1.4948754.

[46] A. Geredakis, M. Karala, N. Ziavra, et E. Toki, « Preliminary measurements of voice parameters using Multi Dimensional Voice Program », World Journal of Research and Review, vol. 5, no 1, 2017.

[47] H. Wu, J. Soraghan, A. Lowit, et G. Di-Caterina, « A Deep Learning Method for Pathological Voice Detection Using Convolutional Deep Belief Networks », in Interspeech 2018, ISCA: ISCA, sept. 2018, p. 446-450. doi: 10.21437/interspeech.2018-1351.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 807-819. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3083

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

SUSTAINABLE STABILIZATION OF LOESS SOILS USING MARBLE POWDER: IMPROVEMENT OF MECHANICAL PROPERTIES

Mohamed Dacher[1], Samir Benmoussa[2], Ouassila Bahloul[3], Mohamed Salah Laouar[4]

[1]Faculty of Technology, Department of Civil Engineering, Laboratory of Civil Engineering – Risks and Structures in Interaction (LGCROI), University of Batna 2, Batna, Algeria.

[2]Laboratory of Numerical Modeling and Instrumentation in Soil-Structure Interaction (NMISSI), Mohamed Khider University – Biskra, Algeria.

[2]Faculdade de Tecnologia, Departamento de Engenharia Civil, Laboratório de Engenharia Civil – Riscos e Estruturas em Interação (LGCROI), Universidade de Batna 2, Batna, Argélia.

[3]Faculty of Science and Technology, Department of Civil Engineering, Mining Laboratory, Echahid Cheikh Larbi Tebessi University, Tebessa, Algeria

[4]Faculty of Science and Technology, Department of Civil Engineering, Mining Laboratory, Echahid Cheikh Larbi Tebessi University, Tebessa, Algeria

[1]https://orcid.org/0009-0008-6871-6802, [2]https://orcid.org/ 0000-0002-9985-238X

[3]https://orcid.org/0000-0001-8554-7563, [4]https:// orcid.org/0000-0003-1767-9075

Email: mohamed.dacher@univ-batna2.dz

ARTICLE INFO

Article History

Received: December 13, 2025

Revised: January 10, 2026

Accepted: January 15, 2026

Published: February 28, 2026

Keywords:

Soil stabilization,
Marble powder,
Settlement potential,
Permeability,
Settable soils.

ABSTRACT

The instability of clayey soils poses a major challenge in geotechnical engineering, especially in arid regions where water infiltration causes a rapid loss of strength and cohesion. Due to the high cost and limited sustainability of conventional stabilization methods, this study evaluates marble powder (MP), a low-cost industrial by-product, as an effective stabilizer for collapsible soils in the Messaad region (Algeria). Soil samples were treated with 5%–20% MP and tested using unconfined compressive strength (UCS), direct shear, California Bearing Ratio (CBR), subsidence tests, and XRD/XRF analyses. The results revealed significant improvements with increasing MP content: UCS increased by 536.36%, CBR by 206.61%, and shear strength by 26.60%. Additionally, maximum dry density increased by 9.14%, while optimum moisture content decreased by 9.86%. XRF results showed higher CaO and lower SiO_2 and Al_2O_3 contents, while XRD confirmed the formation of calcite, which enhances bonding and soil stability. Overall, the study demonstrates that marble powder is an efficient, sustainable, and environmentally friendly material for improving the geotechnical behavior of collapsible soils.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Geotechnical engineering frequently encounters the challenge of unstable soils that threaten the long-term performance and safety of civil structures. In arid and semi-arid regions, two types of problematic soils are particularly prevalent: expansive soils commonly found in Algeria and collapsible loess soils. Expansive soils exhibit volumetric changes due to cycles of wetting and drying, leading to cracking and structural deterioration. In contrast, collapsible soils, such as those in the Djelfa region, experience sudden settlement upon saturation, resulting in a significant loss of bearing capacity [1]. Traditional soil stabilization methods, including the use of lime and cement, are generally effective but often costly, short-lived, and environmentally taxing. The high carbon footprint and resource consumption associated with these materials have prompted increasing interest in more sustainable alternatives [2].

In this context, the reuse of industrial by-products offers a promising pathway toward environmentally responsible soil stabilization. Among these materials, marble powder (MP) a waste product generated by the marble industry stands out for its high calcium carbonate and metal oxide content, which can enhance both the mechanical and hydromechanical behavior of soils [3]. Several recent studies have demonstrated the potential of incorporating marble powder into clayey and collapsible soils. For instance, Nazir et al. (2020) [4] reported that the addition of up to 50% MP reduced the liquid limit by 35%, increased the maximum dry density by 6%, and

improved cohesion by up to 314% under saturated conditions, while also reducing collapse potential [4]. Similarly, [5] observed a substantial reduction in the plasticity and swelling potential of clays, accompanied by enhanced shear strength [5].

More recently, research conducted in China confirmed that soil–marble powder mixtures with up to 60% substitution significantly increased unconfined compressive strength (UCS) and California Bearing Ratio (CBR), while enabling the development of a reliable predictive model for performance [6]. In Algeria, several alternative stabilizers have also been investigated. For example, the treatment of collapsible loess in southern Algeria with sodium silicate reduced collapse potential by more than 87% [7]. Likewise, the use of granulated slag has proven highly effective in improving the shear parameters of subsiding soils [8]. Building on these findings, the present study aims to assess the effectiveness of marble powder as a sustainable stabilizing agent for unstable soils in the Messaad region (Algeria), with particular emphasis on collapsible loess soils from the Djelfa area. Through a comprehensive series of laboratory and in situ tests, this research seeks to develop a sustainable, economical, and environmentally friendly solution that aligns with circular economy principles by valorizing local industrial waste.

II. STUDY AREA AND SAMPLING

The study was conducted in Messaad, located in the Djelfa province of Algeria, approximately 375 km south of Algiers. The region lies at the intersection of the Saharan Atlas and the central steppe (34.09°N, 3.30°E). It is traversed by National Road 89, a strategic 290 km highway connecting M'sila to the southern part of Djelfa. This heavily trafficked route is of particular importance for regional transportation but requires frequent maintenance due to the unstable nature of the underlying soils **Erro! Fonte de referência não encontrada.** Soil samples were collected from a quarry near Messaad (34°18'51"N, 3°37'01" E) Figure 2. A selective sampling approach was adopted to ensure the representativeness of key geotechnical parameters. Disturbed (remanaged) samples were extracted from depths shallower than 1.7 m to minimize organic matter content, while undisturbed samples were obtained using a manual auger for subsequent mechanical testing.

Grain size analysis classified the soil as clay loam (ML) according to the Unified Soil Classification System (USCS). The results are summarized in Table 2 and depicted in the grain size distribution curve (Figure). The investigated section corresponds to the terminal portion of the road, traversing loose clay soils highly susceptible to subsidence. These semi-saturated to dry soils, characteristic of arid environments adjacent to mountainous areas, exhibit strong sensitivity to moisture variations, resulting in settlement and deformation. The objective of this study is to evaluate the susceptibility of these soils to collapse and to propose effective stabilization measures. To achieve this, a comprehensive program of field investigations and laboratory tests was undertaken.

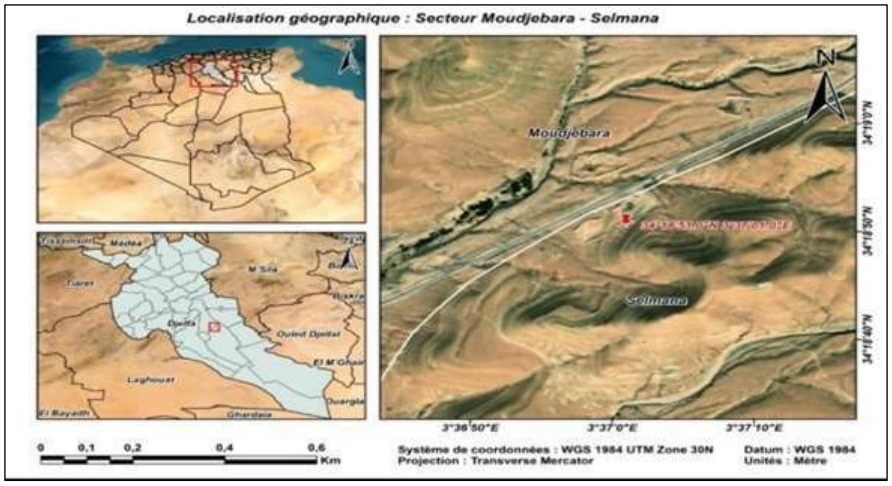


Figure 1: Location of sampling sites.
Source: Authors, (2026).



Figure 2: Natural ravines and collapses in RN89: PK 28+500.

Source: Authors, (2026).

III. MATERIALS AND METHODS

The materials used in this study form the basis of the experimental program. Their prior characterization is essential to assess suitability for soil stabilization. The soil, collected from a quarry along National Route 89 near Messaad (Djelfa province), is classified as clayey silt (ML). Its geotechnical properties were determined in the laboratory before any treatment (Table 1). Marble powder, produced in large quantities during sawing and polishing operations notably from TamStones in Sidi Bel Abbès, the Société Nationale du Marbre in Felfla, and the Chlef marble factory in Aïn Defla is an abundant and inexpensive by-product. Although open-air storage can cause air pollution, its physicochemical characteristics make it a promising material for soil stabilization.

It is composed mainly of calcium carbonate ($CaCO_3$), with a high proportion of fine particles (<75 µm, 70–90%), a specific gravity of 2.65–2.70, and a plasticity index of zero (PI ≈ 0). Its high CaO content (50–56%) contrasts with the low proportion of secondary oxides (SiO_2, Al_2O_3, Fe_2O_3 <3%), moderate MgO (0.5–2%), and alkalis (Na_2O + K_2O <0.5%). Loss on ignition (LOI), linked to decarbonation, ranges between 40 and 45%. These properties confirm that marble powder is chemically stable and has low reactivity, making it effective in improving the geotechnical performance of soils, particularly when combined with hydraulic binders or silica- and alumina-rich additives [9], [10]. Recent studies support its effectiveness; for example, research on coffee husk ash [9] and nanomaterials (Al_2O_3, SiO_2) [10], [11] have shown similar reductions in plasticity and improvements in mechanical stability.



Figure 3: Marble waste in the Messaad (Djelfa) region—Algeria.
Source: Authors, (2026).

III.1 CHARACTERIZATION OF NATURAL SOIL

A series of laboratory tests were conducted to determine the geotechnical properties of the soil before mixing it with marble dust. The results are presented in Table 1.

Table 1: Geotechnical characteristics of loess (RN 89 : PK)268+500).

Désignations	Valeurs
Water content , W(%)	10.4
Specific weight Gs	2.65
Specific gravity of solid particles γ_s(kN/m^3)	27.0
Degree of saturation , Sr (%)	21.07
Saturation content w_{sat} (%)	35.03
Sand (%)	30
Silt (%)	52
Clay (%)	18
Coefficient of uniformity Cu	10.77
Coefficient of curvature Cc	5.49
Maximum dry density γd (g/cm^3)	1.75
Optimum water content w_{opt} (%)	14.2
collapse potentials CP (%)	10.01
Friction angle Φ (degree)	33.16
Cohesion,C(kpa)	11.00

Source: Authors, (2026).

III.2 CHARACTERIZATION OF MARBLE POWDER

The physical and chemical characterization of marble powder (MP) confirms its suitability as a stabilizing additive for fine-grained soils. Its composition is predominantly calcitic, with negligible plasticity, which makes it an effective complementary material in soil stabilization applications.

III.3 CHEMICAL AND MINERALOGICAL ANALYSES

To better understand the influence of marble powder incorporation on soil behavior, detailed chemical and mineralogical analyses were performed using X-ray diffraction (XRD) and X-ray fluorescence (XRF) techniques.

III.3.1 X-Ray Diffraction (XRD)

XRD analysis of the natural soil as shown in the Figure 4. indicates a mineralogical composition dominated by calcite (52.1%), followed by dolomite (43.0%) and a minor proportion of quartz (4.8%). Following the incorporation of marble powder, the diffractograms Figures (5–8) reveal a slight but consistent increase in calcite content, reaching 53.3% for a 5% MP addition (Figure 5), while the proportions of dolomite (42.7–43.1%) and quartz (4.8–5.0%) remain largely stable. At higher MP dosages (10%, 15%, and 20%; Figures 6–8), this trend becomes more pronounced, with progressive intensification of the characteristic calcite peaks. This mineralogical stability, combined with the relative increase in calcite, suggests a gradual enhancement of interparticle bonding. Calcite thus plays a key structuring role by reinforcing intergranular contacts, leading to improved cohesion and mechanical stability in the treated soils.

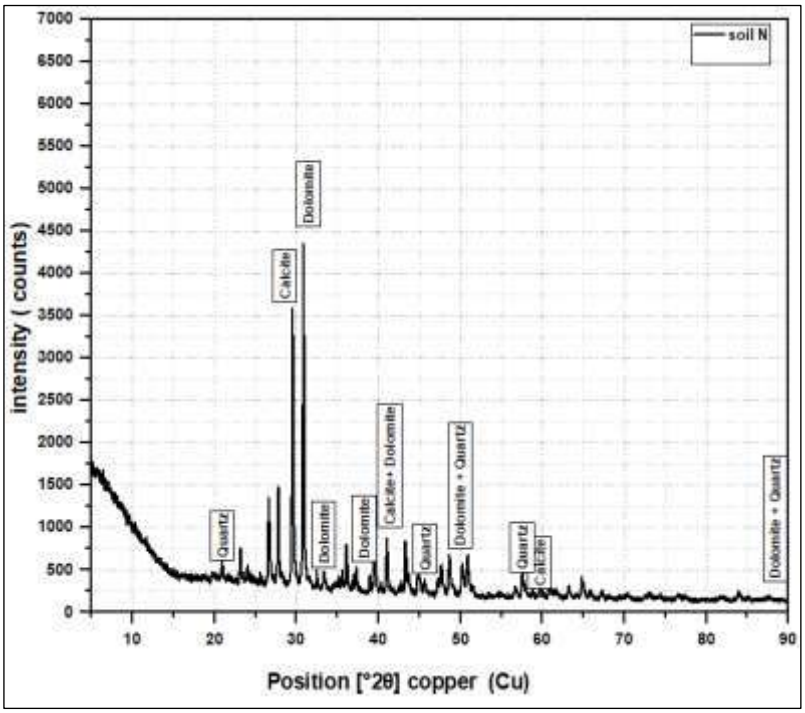


Figure 4: X-ray diffraction (XRD) of natural soil.
Source: Authors, (2026).

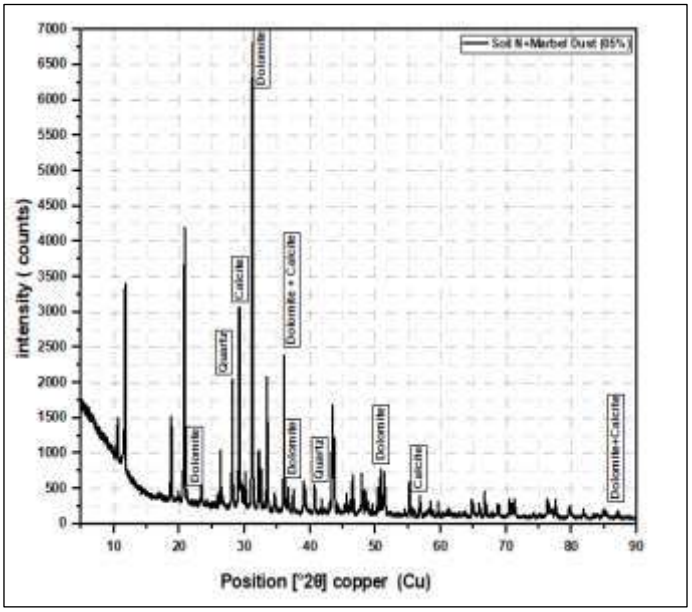


Figure 5: X-ray diffraction (XRD) of the soil treated with 5% marble powder.
Source: Authors, (2026).

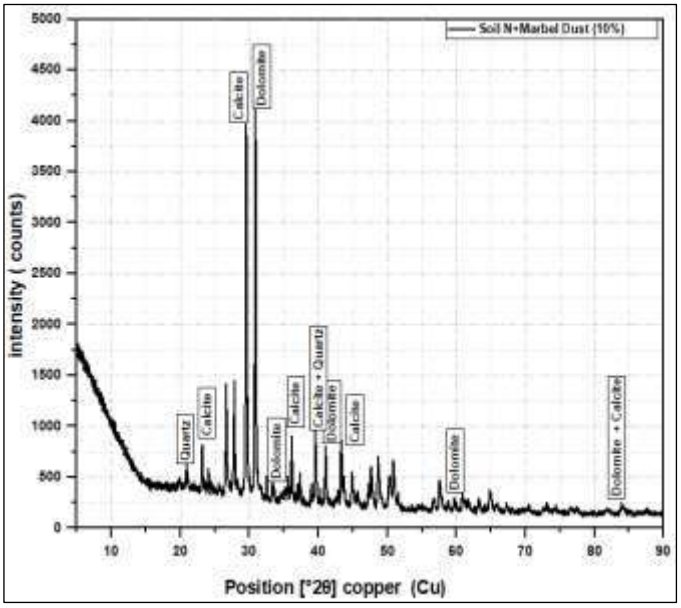


Figure 6: X-ray diffraction (XRD) of the ground treated with 10% marble powder.
Source: Authors, (2026).

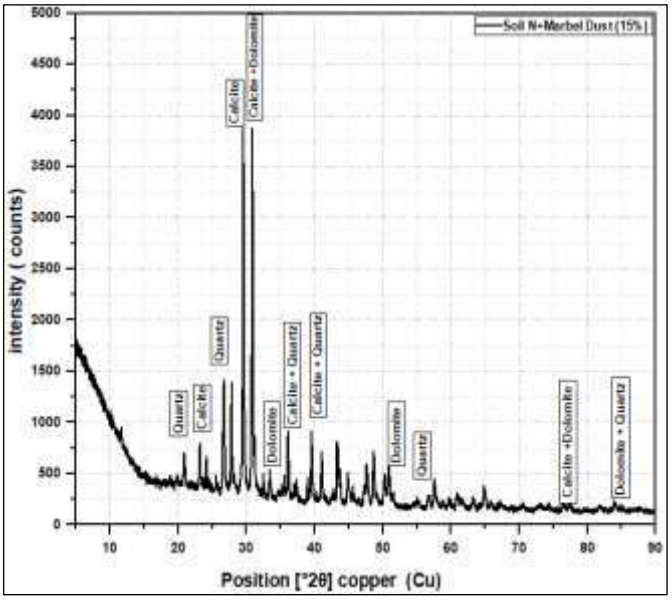


Figure 7: X-ray diffraction (XRD) of the ground treated with 15% marble powder.
Source: Authors, (2026).

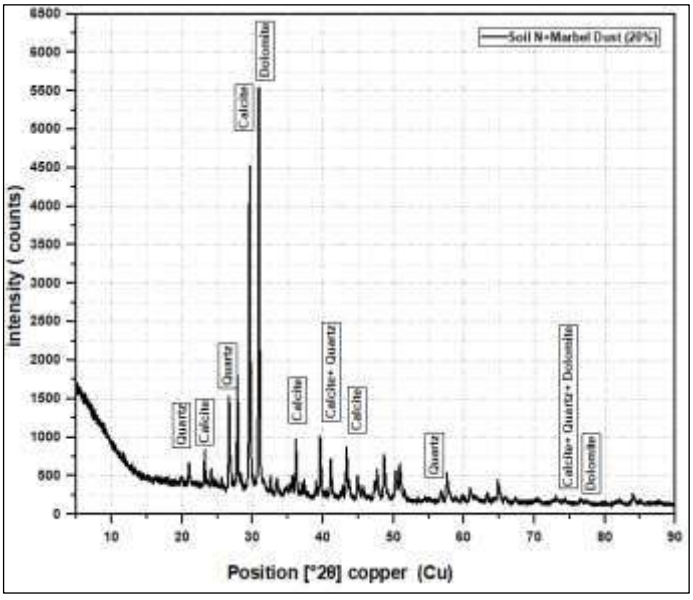


Figure 8: X-ray diffraction (XRD) of the ground treated with 20% marble powder.
Source: Authors, (2026).

III.4 CHEMICAL ANALYSIS BY X-RAY FLUORESCENCE (WDXRF)

The wavelength-dispersive X-ray fluorescence (WDXRF) results (Table *2*) reveal a marked increase in calcium oxide (CaO) content from 34.87% in the natural soil to 39.74% after the addition of 10% marble powder confirming the contribution of calcium carbonate and the activation of cementation mechanisms [1], [2]. Concurrently, the concentrations of silicate oxides exhibit a gradual decrease (SiO_2: 11.54% → 9.46%; Al_2O_3: 4.36% → 3.36%), indicating reduced susceptibility to moisture variations and improved structural stability [3].

In addition, the contents of sulfur trioxide (SO_3) and chloride ions (Cl^-) decrease by approximately 18% and 32%, respectively, reflecting a reduction in soluble salts that are often associated with geotechnical instability [4]. These compositional changes confirm that the incorporation of marble powder promotes the formation of stable carbonate phases, which enhance soil cohesion, reduce collapse potential, and improve overall mechanical performance. These findings are consistent with recent research on clayey soils treated under varying confinement conditions [12], which similarly demonstrated that fine mineral additives mitigate water sensitivity and enhance mechanical strength.

Table 2: XRF chemical composition of soil samples with varying marble dust contents.

	SOIL N T	SN- M 5%	SN- M 10%	SN- M 15%	SN- M 20%
Na_2O	0.8094	0	0	0	0
MgO	5.0029	4.6643	4.1149	5.1237	3.9825
Al_2O_3	4.3606	4.0785	3.827	4.1425	3.3619
SiO_2	11.546	11.317	10.722	10.701	9.4598
P_2O_5	0.0727	0.0858	0.0806	0.0716	0.0663
SO_3	0.7258	0.7005	0.6643	0.6482	0.591
Cl	0.2037	0.1803	0.1743	0.1707	0.1379
K_2O	1.7052	1.7243	1.6899	1.5712	1.4873
CaO	34.877	37.542	39.738	36.997	38.992
TiO2	0.1325	0.127	0.2027	0	0.0993
MnO	0.0188	0.0223	0.0278	0.0183	0.023
Fe_2O_3	1.0992	1.2271	1.3133	1.3977	1.152
ZnO	0	0.0043	0	0	0
Br	0	0.0028	0.0043	0.0037	0.0028
Rb_2O	0	0.0031	0.0049	0.0037	0.0038
SrO	0.0488	0.0561	0.0598	0.0657	0.0527
ZrO_2	0.0024	0.0039	0.0046	0	0
CO_2	40.62	39.25	38.09	37.5	39.24

Source: Authors, (2026).

IV. EXPERIMENTAL METHODS

IV.1 PREPARATION OF SOIL–MARBLE POWDER SAMPLES

The marble powder (MP) percentages selected for this study—5%, 10%, 15%, and 20%—were chosen based on previous research demonstrating that this range is particularly effective for improving the geotechnical properties of fine soils stabilized with industrial residues. [11] reported that incorporating 5-20% fine additives into expansive soils results in substantial gains in both mechanical strength and overall stability. Similarly, several recent studies have confirmed that increasing the proportion of marble powder significantly decreases the plasticity index and swelling potential of clayey soils, thereby enhancing their suitability for road construction and shallow foundations [9], [10], [12] found that a 15% MP dosage yielded the most favorable California Bearing Ratio (CBR) value (~8%), accompanied by a marked reduction in plasticity an observation that supports the dosage levels adopted in this study.

At higher concentrations, Umar and Lin (2024) [6] observed a considerable improvement in bearing capacity, with the CBR increasing from approximately 10.4% to 22.9% and the unconfined compressive strength (UCS) rising from 170 to 661 kN/m² after 28 days of curing. These results confirm the effectiveness of marble powder in enhancing mechanical performance. Boukhatem et al. (2024) [13] further demonstrated that combining marble powder with lime reduces the void ratio and increases soil density while promoting certain pozzolanic reactions. Similarly, [14], [15] reported that the use of calcium carbide residue improves the mechanical properties of fine soils through stabilization mechanisms comparable to those achieved with marble powder. In a related study. [16], [17] showed that combining lime with digestates derived from rice waste provides a sustainable and efficient method for clay soil stabilization, supporting the broader use of industrial and agro-industrial by-products.

X-ray diffraction (XRD) analysis of the marble powder revealed a crystalline structure dominated by calcite ($CaCO_3$), with intense peaks around 29.4° 2θ, characteristic of this carbonate mineral. Secondary reflections observed between 23° and 48° 2θ confirmed the presence of dolomite ($CaMg(CO_3)_2$) and, to a lesser extent, quartz (SiO_2). These stable and inert mineral phases indicate that marble powder primarily acts as a mineral filler rather than as a pozzolanic material. Scanning electron microscopy (SEM) analyses showed angular particles of varying sizes with sharp edges, promoting better interlocking with soil grains. Elemental mapping confirmed high calcium and magnesium contents consistent with the carbonate nature of the material. This morphology, combined with the fine particle size, reduces intergranular voids and increases the maximum dry density during compaction.

Although marble powder does not exhibit strong pozzolanic activity, its high carbonate content and favorable physical characteristics make it an effective additive for physical soil stabilization mainly through densification, improved compactness, and reduced plasticity. These findings are consistent with recent research on the use of industrial residues and nanomaterials in the treatment of fine soils, which highlights not only significant gains in strength and improved durability of stabilized mixtures [9-12], Umar & Lin (2024) [6], Boukhatem et al. (2024) [13], but also the emergence of new sustainable approaches incorporating industrial and agricultural by-products [14], [16], Hanafi et al. (2025) [18].

IV.2 EXPERIMENTAL PROGRAM AND ASSESSMENT OF SUBSIDENCE POTENTIAL

A comprehensive experimental program was conducted to evaluate the influence of marble powder on the geotechnical behavior of the studied soils, with particular emphasis on their susceptibility to subsidence. The first stage involved granulometric analysis of the natural soil, which is essential for identifying the open, loose structure characteristic of subsidence-prone soils particularly loess typically distinguished by high porosity and low dry density [15], [17]. The second stage focused on quantifying the subsidence potential using a subsidence susceptibility index derived from single and double oedometer tests [19], [20], conducted in accordance with ASTM D5333-03 [21], [22]. This procedure measures volumetric deformation in response to progressive saturation, providing an accurate assessment of soil reactivity during water infiltration [23]. Together, these two components grain-size characterization and evaluation of collapse susceptibility constitute a robust methodological framework. They enable a rigorous comparison of soil properties before and after marble powder treatment, facilitating the assessment of both microstructural and mechanical improvements.

IV.3 GRAIN SIZE DISTRIBUTION OF NATURAL SOIL

Grain size analysis of the loess samples Figure shows a predominance of silt (52%), followed by sand (30%) and clay (18%). The uniformity coefficient (Cu = 10.77), derived from the grain size curve, indicates a high susceptibility to subsidence (Ayadat and Belhouari). A second value (Cu = 10.77) further confirms the presence of a transitional interval conducive to collapse phenomena. These results characterize the natural soil's behavior and highlight its pronounced susceptibility to subsidence. To assess potential improvements and mitigate this risk, a subsequent series of analyses was conducted on soil stabilized with marble powder.

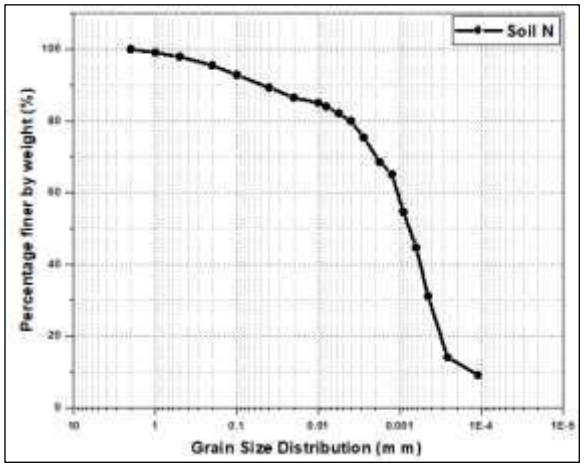


Figure 9: Grain size distributions curve for tested soil.
Source: Authors, (2026).

IV.4 ASSESSMENT OF SUBSIDENCE POTENTIAL

The subsidence potential of the natural soil was evaluated using oedometer tests, which quantify the tendency of the material to undergo sudden compression upon water infiltration. The settlement index obtained serves as a key indicator of the soil's susceptibility to structural instability. According to the criteria proposed by [24], soil is classified as non-settling when the uniformity coefficient (Cu) is ≤4%, transitional when 4% < Cu < 12%, and clearly settling when Cu ≥ 12%. For the soil studied (RN89, PK 268+500), the value of Cu=10.77% places it firmly in the category of subsidence-prone soils, confirming its high vulnerability to saturation. These findings align with several recent studies, which report that loess soils and metastable wind-blown deposits are highly prone to collapse when moistened [22], [24–26].

Furthermore, experimental and numerical investigations have validated the use of the subsidence index as an effective predictive tool for unstable behavior, while emphasizing its integration into multi-criteria approaches that consider soil particle size, suction, and microstructure [27], [28]. The results obtained in this study corroborate these observations and underscore the necessity of stabilization treatments such as the incorporation of marble powder to reduce subsidence susceptibility and enhance the geotechnical durability of the soils under investigation.

IV.5 EVALUATION OF COMPRESSIBILITY TESTS

Compressibility tests performed on the natural soil allowed the calculation of the collapsibility index (Cp), a widely used parameter for quantifying the susceptibility of loess soils to subsidence. According to the classification of Jennings and Knight (1975), Cp values are categorized from "non-subsiding" to "very severely subsiding" soils, depending on the magnitude of settlement under applied stress. For the soil studied, a vertical load of σ = 200 kPa resulted in Cp = 10.01%, placing it in the "troubled" category. This indicates a significant potential for subsidence, which could compromise the stability of foundations. These findings are consistent with recent research confirming the relevance of the Cp index as a diagnostic tool for loess soils, particularly in semi-arid regions where the open soil structure is highly vulnerable to water infiltration [29–33]. The combined use of particle size analysis, Atterberg limits, and compressibility testing provides a robust framework for assessing subsidence behavior. This evaluation is an essential preliminary step

IV.6 EVALUATION OF DIRECT SHEAR STRENGTH

The direct shear test conducted on the natural soil indicated very low cohesion (c = 11 kpa)) and a moderate angle of internal friction (φ = 33.16°). These results suggest that the soil's overall strength is primarily governed by friction, with cohesion contributing minimally to its mechanical behavior. This pattern is typical of loess soils, whose open structure, low active clay content, and sandy-to-silty grain size distribution lead to reduced cohesion and internal friction generally ranging between 30° and 35°. Several studies have highlighted this behavior.

For example, [24], [33] reported that loess soils in Algeria exhibit negligible cohesion but relatively high internal friction, reflecting their fragile structure while maintaining granulometric properties conducive to frictional resistance. Recent research further confirms that the shear strength of loess soils is strongly influenced by factors such as degree of saturation and initial density [30], [34], [35]. Therefore, the relatively high friction angle observed in this study aligns with the silty and slightly clayey nature of the soil, but it does not offset the soil's susceptibility to subsidence when saturated.

IV.7 GENERAL METHODOLOGY

The methodology followed in this study, illustrated in Figure 10, involves the preparation and characterization of the materials—subsiding soils from Djelfa and marble powder derived from industrial waste—followed by the preparation of soil–marble powder mixtures with varying additive ratios (0–20%). Standardized laboratory tests were then conducted to evaluate the behavior of these mixtures. The investigations focused sequentially on: Compaction properties: including modified Proctor tests, maximum dry density, and optimum moisture content. Mechanical performance: assessed through unconfined compressive strength (UCS), California Bearing Ratio (CBR), and direct shear tests. Hydromechanical behavior: evaluated via subsidence susceptibility measurements.

This integrated approach allows for a systematic assessment of the influence of marble powder on the geotechnical properties of the studied soils and facilitates the identification of optimal proportions for sustainable stabilization.

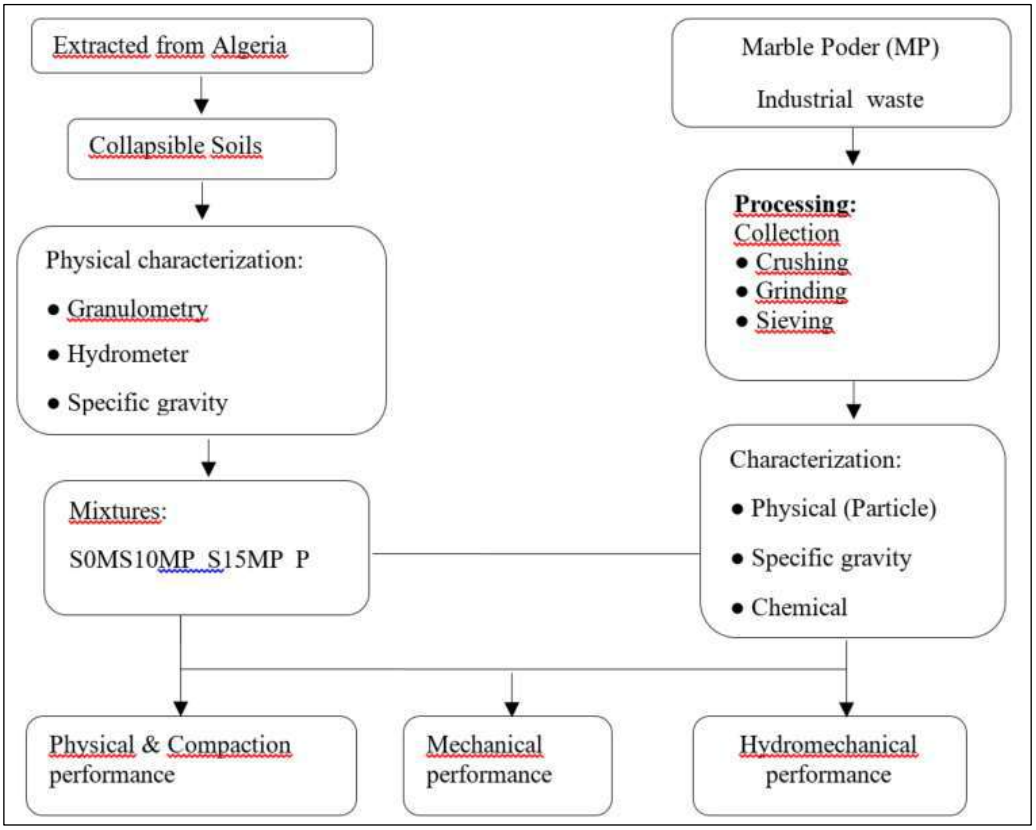


Figure 10: Experimental methodology for stabilizing collapsible soil with marble powder.
Source: Authors, (2026).

V. RESULTS AND DISCUSSIONS

This section aims to analyze and discuss the results of the experiments in order to highlight the effect of adding marble powder on the geotechnical properties of the soil under study, through a methodological comparison between the behavior of natural soil and that of stabilized mixtures. The discussion begins with a presentation of the results of the chemical analysis of both the soil and the marble powder, with the aim of explaining the changes in mechanical properties depending on the chemical composition of the materials used, particularly the presence of calcium-rich compounds in the marble powder.

Next, the effect of the addition on the mechanical properties, especially compressive strength, is examined, in addition to other parameters such as maximum dry density, optimum moisture content, and deformation properties. This integrated approach analysis to link mechanical results to chemical background, contributing to a deeper understanding of the mechanism of action of marble powder and evaluating its effectiveness as a soil conditioner from a scientific and practical standpoint... Chemical and mineralogical analyses To better understand the effect of adding marble powder on soil behavior, detailed chemical and mineralogical analyses were performed using X-ray diffraction (XRD) and X-ray fluorescence (XRF) techniques.

V.1 EFFECT OF MARBLE POWDER ON COMPACTION PROPERTIES

The influence of varying proportions of marble powder (MP) on soil compaction properties was evaluated using modified Proctor tests in accordance with ASTM D698. Mixtures containing 0%, 5%, 10%, 15%, and 20% MP were prepared, and the results were plotted Figures (11–13) to illustrate the relationship between water content and dry density. The results indicate that increasing the marble powder content leads to a significant rise in maximum dry density compared to the unsterilized soil, which had a value of 1.75 g/cm³. Specifically, maximum dry density increased by 5.71%, 6.86%, 8%, and 9.14% for MP contents of 5%, 10%, 15%, and 20%, respectively, with the highest value reaching 1.91 g/cm³. This improvement is primarily due to the filling effect, whereby fine marble powder particles occupy the voids between larger soil grains, reducing porosity and enhancing overall compactness [1], [25].

In addition, the higher specific gravity of marble powder, combined with particle rearrangement and flocculation, contributes to improved densification and cohesion in the stabilized soil. Conversely, the optimum moisture content (OMC) exhibited a decreasing trend with increasing MP content. The reductions relative to the natural soil were 2.11%, 0.7%, 0.7%, and 9.86% for 5%, 10%, 15%, and 20% MP, respectively. This decrease is attributed to the low plasticity and low water absorption of marble powder, which reduces the amount of water required for effective particle lubrication [6], [25]. Additionally, the increased specific surface area of the mixture promotes more uniform water distribution, further lowering the OMC. These findings are consistent with previous studies, including those by [36-38], which reported that marble powder incorporation enhances maximum dry density, reduces plasticity, and favorably influences the compaction behavior of fine soils.

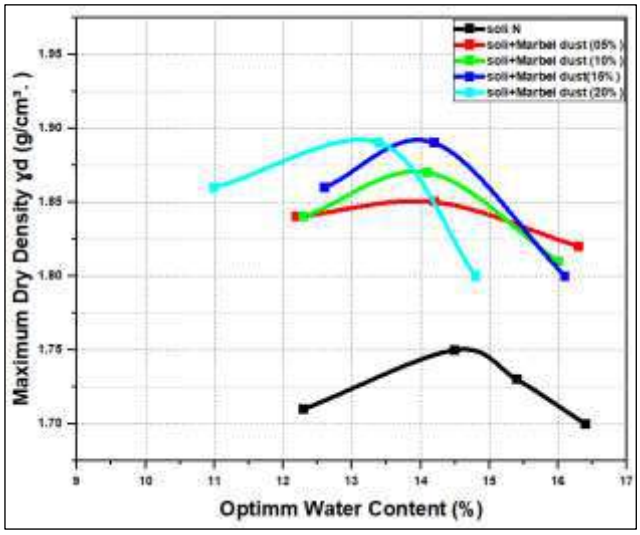


Figure 1: Influence of marble dust on Optimum parameters of proctor test.
Source: Authors, (2026).

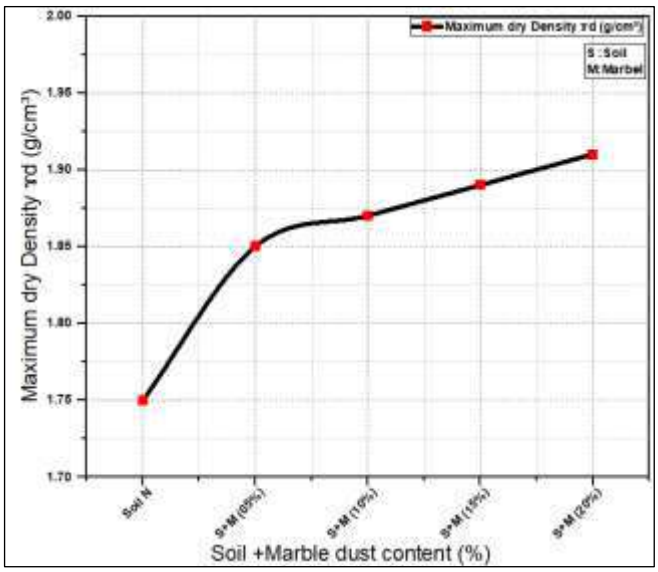


Figure 2: Influence of marble dust on maximum dry density.
Source: Authors, (2026).

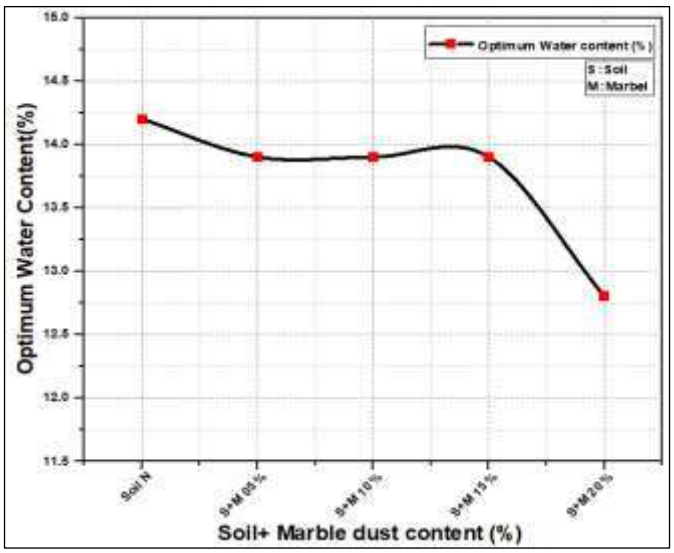


Figure 3: Influence of marble dust on Optimum Water Content (%).
Source: Authors, (2026).

V.2 EFFECT OF MARBLE POWDER ON THE CALIFORNIA BEARING RATIO (CBR) AND DRY DENSITY OF SOIL

To evaluate the behavior of stabilized soil under water exposure, the effect of varying marble powder (MP) proportions (0%, 5%, 10%, 15%, and 20%) on the California Bearing Ratio (CBR) and dry density (γd) was investigated. CBR tests were conducted on samples immersed for 96 hours to simulate prolonged water exposure, such as during sustained rainfall or subsoil infiltration. This saturation period allows the soil to fully absorb water, revealing its potential for settlement and strength reduction, while ensuring comparability with standard recommendations and literature data. CBR values were determined at 95% of the maximum dry density obtained from modified Proctor tests. The results are presented in Figures 14 and 15. The data show a clear positive correlation between marble powder content and soil CBR values.

Specifically, increasing MP content from 0% to 20% raises the CBR from 5.14% to 15.76%. This improvement is attributed to several mechanisms: Void Filling Effect: Fine marble particles occupy the interstitial spaces between larger soil grains, increasing soil density and reducing total void volume. This limits water absorption and enhances post-immersion stability and strength, as reflected by the corresponding increase in dry density (γd) with higher MP content. Minimum Acceptable Threshold Achieved: With the addition of 10% MP, the CBR value (10.07%) surpasses the typical minimum requirement for road sub-bases (approximately 8%). This indicates that an appropriate proportion of marble powder can effectively improve the bearing capacity of weak soils, making them suitable for pavement layers.

Increase in Dry Density: The gradual rise in dry density with increasing MP content contributes directly to higher CBR values, as denser soils generally exhibit greater resistance to deformation. Reduction in Expansion Upon Immersion: Although specific expansion measurements were not recorded, the 96-hour immersion CBR test reflects the impact of water absorption on soil durability. Filling voids with marble powder likely reduces permeability and limits expansion, helping maintain higher durability after water exposure. In summary, the addition of marble powder simultaneously enhances soil density and bearing capacity after immersion, providing an effective solution for strengthening weak soils intended for pavement applications.

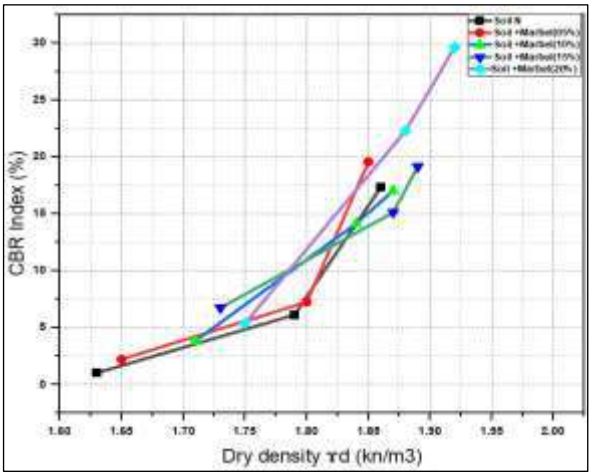


Figure 4: Influence of Maximum Dry Density (MDD) on CBR Index.
Source: Authors, (2026).

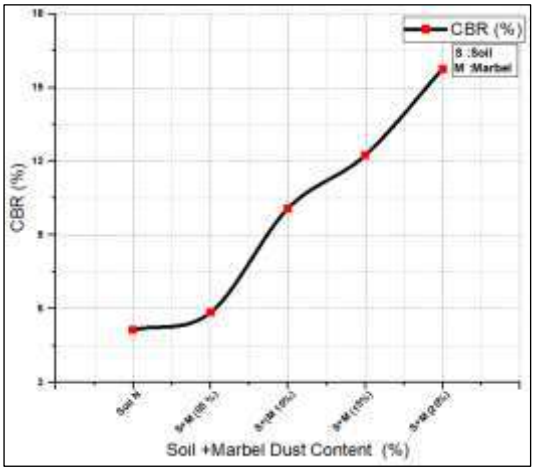


Figure 5: Influence of marble dust on CBR index (%).
Source: Authors, (2026).

V.3 EFFECT OF MARBLE POWDER ON THE DIRECT SHEAR PROPERTIES OF SOIL

Direct shear tests were conducted on both treated and untreated soil samples. The treated samples were prepared with varying marble powder (MP) contents of 5%, 10%, 15%, and 20% by dry weight, while untreated control samples (0% MP) were also tested. Different vertical stresses were applied to the samples, and the maximum shear stress was recorded to establish the Mohr-Coulomb failure envelope. From this envelope, the internal friction angle (φ) and cohesion (c) were determined for each mixture.

V.3.1 Internal friction angle (φ)

The results Figure *6* show a slight decrease in the internal friction angle with increasing marble powder content. The measured values are 32.12°, 32.12°, 32.18°, 29.17°, and 29.12°, compared to 33.16° for the untreated soil. In relative terms, these represent reductions of approximately 3.1%, 3.1%, 3.0%, 12.0%, and 12.2%, respectively. This decrease can be attributed to the modification of soil grain surfaces caused by the introduction of fine, relatively smooth marble powder particles. These particles reduce interlocking and friction between coarser grains, resulting in slightly lower resistance to sliding along the shear plane. However, the reduction remains minor across the tested proportions, indicating that marble powder has a negligible effect on the internal friction angle [18], [26].

V.3.2 Cohesion (c)

In contrast, a significant increase in soil cohesion was observed with higher marble powder contents Figure *6*: Effect of marble dust content on shear strength parameters (internal friction angle ϕ and cohesion c). Cohesion values increased from 11 kPa in the untreated soil to 17, 48, 62, and 70 kPa for 5%, 10%, 15%, and 20% MP, respectively, corresponding to relative increases of approximately +55%, +336%, +464%, and +536%. Cohesion, representing the soil's shear strength in the absence of vertical stress, is therefore greatly enhanced by marble powder addition. This increase reflects the strengthening of intermolecular forces in the soil, which can be attributed to two main mechanisms: (i) Filling of Interstitial Voids: Fine marble powder particles occupy void spaces, reducing porosity and increasing grain contact surfaces. (ii) Bridge Formation Between Grains: Marble particles can form bonds between soil grains, improving shear strength under low stress and increasing apparent cohesion. These results indicate that, unlike the internal friction angle, which remains relatively stable, marble powder substantially increases cohesion, thereby enhancing the overall mechanical strength of the soil.

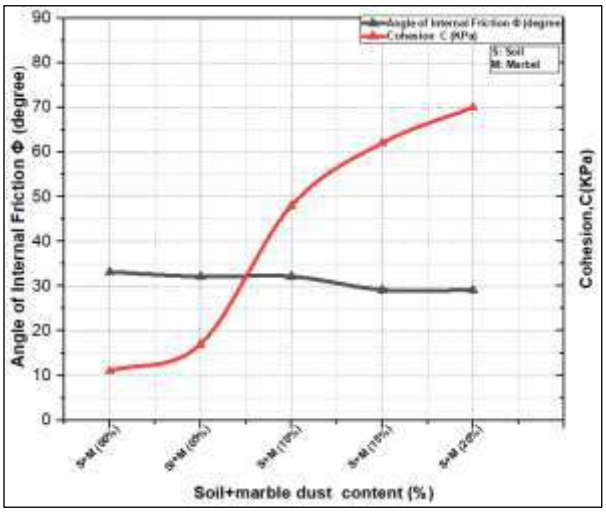


Figure 6: Effect of marble dust content on shear strength parameters (internal friction angle ϕ and cohesion c).
Source: Authors, (2026).

V.4 EFFECT OF MARBLE POWDER TREATMENT ON SLUMP POTENTIAL

The effect of marble powder (MP) addition on the slump potential (CP) of collapsible soils was evaluated. The untreated soil exhibited unstable behavior when immersed in water, with a CP value of 10.01%, classifying it as "beginning of severe problems" according to the Jennings and Knight scale (CP > 5) [9], [39]. This study aimed to determine the effectiveness of marble powder in improving soil stability and reducing susceptibility to settlement. Four soil samples treated with 5%, 10%, 15%, and 20% MP were tested using standard oedometer tests under a constant vertical stress of 200 kPa. The CP values and their reduction relative to the untreated soil are presented in Figures 17 and 18: (i) 5% MP: CP decreased to 6.18%, a reduction of 37.92%.

Despite this improvement, the soil remains classified as "beginning of severe problems" (CP = 6.18 > 5). (ii) 10% MP: CP dropped to 2.05%, a 79.24% reduction, improving the classification to "moderate problems" ($1 < CP \leq 5$). (iii) 15% MP: CP further decreased slightly to 2.01%, a 79.60% reduction, maintaining the "moderate problems" classification. (iv) 20% MP: CP reached the lowest value of 1.42%, an 85.01% reduction, with the soil still classified as "moderate problems". These results demonstrate that increasing marble powder content progressively reduces the collapsibility of subsiding soils. This improvement can be attributed to several mechanisms: the filling of voids by fine marble particles, enhanced internal friction between grains, and minor particle interactions that contribute to increased cohesion [13], [14], [15].

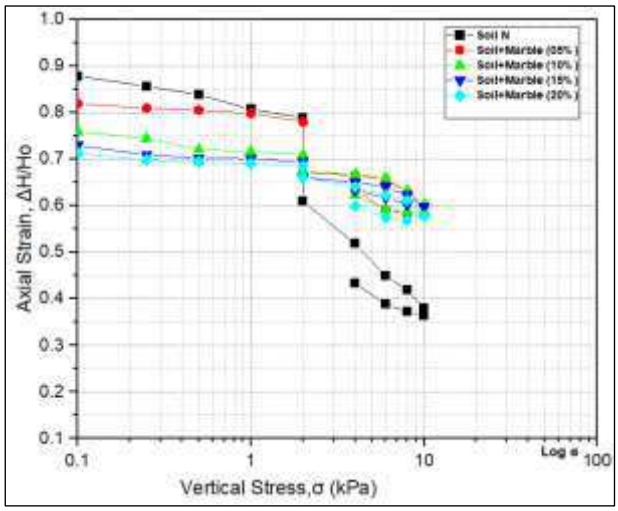


Figure 7: Collapse potential curves for marble dust – soil mixture.
Source: Authors, (2026).

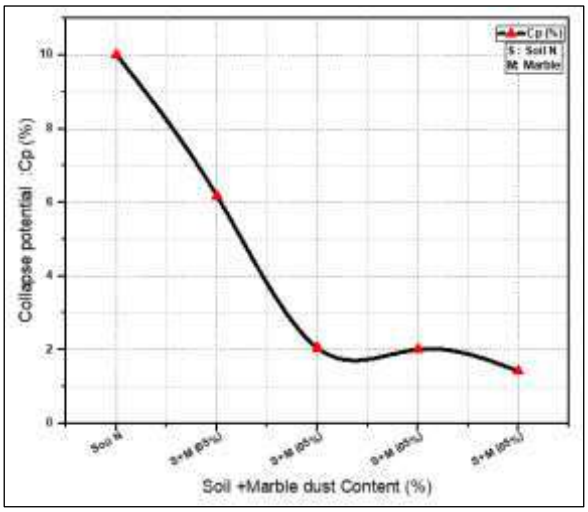


Figure 8: Influence of marble dust on collapse potential Cp (%).
Source: Authors, (2026).

VI. CONCLUSIONS

The results of this study confirm the effectiveness of marble powder (MP) as an innovative and sustainable stabilizer for improving the physical and mechanical properties of subsiding soils. The main findings are as follows:

✓ Improved Compaction Properties: Addition of up to 20% MP increased the maximum dry density and decreased the optimum moisture content, indicating enhanced compaction behavior.

✓ Strength and Bearing Capacity Enhancement: At 15% MP, the soil exhibited higher cohesion and resistance. The California Bearing Ratio (CBR) increased significantly, demonstrating that treated soils can be effectively used for road foundation layers.

✓ Reduced Settlement Potential: Oedometer tests showed a decrease in settlement potential with increasing MP content, attributed to void filling and higher density, which improves resistance to subsidence.

✓ Optimization of Dosage: Results indicate that 10% MP may suffice to meet minimum road sub-base requirements. Emphasizing however, that the optimal dosage should be determined based on economic considerations, soil characteristics, and project specifications.

✓ Overall Improvement in Stability: MP treatment significantly reduces the susceptibility of subsiding soils, improves hydromechanical behavior, and limits the risk of differential settlement.

✓ Reduced Permeability and Improved Durability: Permeability decreased progressively with increasing MP content, reaching a reduction of over 90% at 20% MP. This limits water infiltration, reduces subsidence risk, and enhances long-term durability for civil engineering applications.

✓ Enhanced Soil Cohesion: Cohesion increased substantially, from 11 kPa in untreated soil to 70 kPa with 20% MP, improving shear strength and overall stability.

✓ Sustainable and Environmentally Friendly Solution: The use of marble powder valorizes an abundant industrial by-product while minimizing environmental impact.

✓ Improved Resistance to Moisture-Drying Cycles: Incorporation of MP also helps mitigate cracking and differential settlement caused by repeated wetting and drying.

Outlook:

✓ For future applications, it would be relevant to study the combined effect of marble powder with other binders (such as cement or lime) to further optimize soil stability and strength.

Additional studies on long-term behavior and cyclic loading could complement this approach and ensure the durability of structures built on these treated soils.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Mohamed Dacher, Samir Benmoussa, Ouassila Bahloul, Mohamed Salah Laoar.
Methodology: Mohamed Dacher, Samir Benmoussa.
Investigation: Mohamed Dacher, Samir Benmoussa, Ouassila Bahloul.
Discussion of results: Mohamed Dacher, Samir Benmoussa, Ouassila Bahloul, Mohamed Salah Laoar.
Writing – Original Draft: Mohamed Dacher, Samir Benmoussa.
Writing – Review and Editing: Mohamed Dacher, Samir Benmoussa, Ouassila Bahloul.
Resources: Mohamed Dacher, Ouassila Bahloul.
Supervision: Samir Benmoussa.
Approval of the final text: Mohamed Dacher, Samir Benmoussa.

VIII. ACKNOWLEDGMENTS

The author wishes to express sincere gratitude to Mr. Tawfiq Zekhmi, colleague at the University of Batna 2, for his valuable help and constructive support throughout the development of this research.

IX. REFERENCES

[1] D. B. Hamdy, 'A sustainable approach for improving the behavior of collapsible soil using marble dust', Innov. Infrastruct. Solut., vol. 7, no. 4, p. 260, Aug. 2022, doi: 10.1007/s41062-022-00859-z.

[2] M. Elsiragy, 'Geotechnical behaviour of reinforced soft clay by marble dust as a waste material', Journal of Engineering Research, vol. 0, pp. 0–0, 2021, doi: 10.21608/erjeng.2021.88123.1023.

[3] A. Debnath, S. Saha, and R. Chattaraj, 'Stabilization of clayey soil with marble dust', in Recent Developments in Sustainable Infrastructure, vol. 75, B. B. Das, S. Barbhuiya, R. Gupta, and P. Saha, Eds, Singapore: Springer, 2021, pp. 175–183. doi: 10.1007/978-981-15-4577-1_14.

[4] A. Nazir, M. E. El Sawwaf, W. Azzam, and M. Ata, 'Utilization of marble dust for improving the geotechnic characteristics of collapsible soil', Journal of Geological Research, vol. 2, no. 4, pp. 36–45, 2020, doi: 10.30564/jgr.v2i4.2352.

[5] A. Saygılı, 'Use of waste marble dust for stabilization of clayey soil', Materials Science (Medžiagotyra), vol. 21, no. 4, pp. 600–604, 2015, doi: 10.5755/j01.ms.21.4.11966.

[6] I. H. Umar and H. Lin, 'Marble powder as a soil stabilizer: An experimental investigation of the geotechnical properties and unconfined compressive strength analysis', Materials, vol. 17, no. 5, p. 1208, 2024, doi: 10.3390/ma17051208.

[7] O. Bahloul, F. Z. Tebbi, and L. Lekouara, 'Experimental study of the physico-chemical stabilization of a loess by sodium silicate: Case of Southern Algeria', Selected Scientific Papers – Journal of Civil Engineering, vol. 18, no. 1, 2023, doi: 10.2478/sspjce-2023-0013.

[8] H. Ziani, S. Deboucha, and A. Amriou, 'Influence of granulated slag on the mechanical parameters of compacted collapsible soil', in Proceedings of the 18th African Regional Conference on Soil Mechanics and Geotechnical Engineering (ARCSMGE), 2024.

[9] M. Riascos-Caipe, V. Mora-Ruiz, F. Nieto-Castañeda, A. Hernández-García, and M. Herrera-Vargas, 'Evaluation of clay soil stabilization with coffee husk ash for sustainable geotechnical applications', Revista Matéria, 2025, doi: 10.1590/1517-7076-RMAT-2025-0049.

[10] M. R. Abisha and J. P. A. Jose, 'Experimental investigation on soil stabilization technique by adding nano-aluminium oxide additive in clay soil', Revista Matéria, 2022, doi: 10.1590/1517-7076-RMAT-2022-0272.

[11] M. R. Abisha, 'Experimental analysis of clay soil with nanosilicon dioxide', Revista Matéria, 2024, doi: 10.1590/1517-7076-RMAT-2024-0858.

[12] R. P. Aparna, R. G. Robinson, and S. R. Gandhi, 'Experimental analysis on the effect of overburden stress on the smear zone characteristics of clayey soil', Soils and Rocks, 2025, doi: 10.28927/SR.2025.005524.

[13] G. Boukhatem and others, 'Optimizing properties of clayey soil using lime and waste marble powder', Frontiers in Materials, 2024.

[14] P. Tang and others, 'Calcium carbide residue for clay stabilisation: mechanical performance & mechanism', 2025.

[15] K. Abbeche, T. Ayadat, and A. Lahmadi, 'Traitement d'un sol á effondrement brusque par la chaux lime stabilisation of a collapsible soil', Séminaire International Innovation et Valorisation dans le Génie Civil, Hammamet, Tunisia, pp. 161–168, 2009.

[16] A. Sendilvadivelu and others, 'Sustainable Stabilization of Clay Soil Using Lime and Oryza sativa-Waste-Derived Digestate', Sustainability, 2025.

[17] T. Ayadat, 'Note technique: Traitement d'un sol affaissable par la chaux', Revue française de géotechnique, no. 77, pp. 69–72, 1996.

[18] M. Hanafi and others, 'Evaluating the strength, durability and porosity of soils improved with marble dust', 2025.

[19] T. O. Al-Refeai and A. A. Al-Karni, 'Experimental Study on the Utilization of Cement Kiln Dust for Ground Modification', Journal of King Saud University - Engineering Sciences, vol. 11, no. 2, pp. 217–231, 1999, doi: 10.1016/S1018-3639(18)30999-1.

[20] T. Singh, P. Pattnaik, D. Shekhawat, L. Ranakoti, and L. Lendvai, 'Waste marble dust-filled sustainable polymer composite selection using a multi-criteria decision-making technique', Arabian Journal of Chemistry, vol. 16, no. 6, p. 104695, June 2023, doi: 10.1016/j.arabjc.2023.104695.

[21] ASTM International, Standard Test Method for Measurement of Collapse Potential of Soils (ASTM D5333-03). West Conshohocken, PA, 2018.

[22] ISSMGE, 'Evaluation of collapsible soils in arid regions', in Proceedings of ARCSMGE, 2024, pp. 1–8.

[23] O. Bahloul, H. Ziani, and S. Benmoussa, 'Impact of calcium chloride on the microstructure of a collapsible soil', Annales de Chimie Science des Matériaux, vol. 46, pp. 201–206, 2022, doi: 10.18280/acsm.460405.

[24] T. Ayadat and A. Belhouari, 'Engineering properties of collapsible soils in Algeria', Geotechnical and Geological Engineering, vol. 19, no. 3–4, pp. 189–205, 2001.

[25] M. Y. Fattah and others, 'Evaluation of collapse potential in different collapsible soils', Geoderma, vol. 425, p. 116052, 2022.

[26] T. H. Bui, 'Collapse behavior of loess soils: Laboratory study and numerical modeling', Soils and Foundations, vol. 63, no. 1, pp. 101–115, 2023, doi: 10.1016/j.sandf.2022.11.006.

[27] A. Khaldoun, 'Assessment of collapsible soil behavior using oedometer and suction-based approaches', Engineering Geology, vol. 293, p. 106120, 2021, doi: 10.1016/j.enggeo.2021.106120.

[28] Y. Zhang, 'Microstructural evolution and collapse potential of loess under wetting conditions', Applied Clay Science, vol. 241, p. 106915, 2023, doi: 10.1016/j.clay.2023.106915.

[29] T. Ayadat and W. Wally, 'Collapse potential of compacted soils', Proceedings of the Institution of Civil Engineers – Geotechnical Engineering, vol. 143, no. 1, pp. 45–52, 2000.

[30] S. L. Houston, C. E. Zapata, and C. Lawrence, 'Characterisation and assessment of collapsible soils', Soils and Foundations, vol. 61, no. 5, pp. 1181–1198, 2021.

[31] J. Feda, Mechanics of Soils in Collapse Compressibility. Amsterdam: Elsevier, 1996.

[32] H. Nowamooz and F. Masrouri, 'Hydromechanical behaviour of collapsible soils: Characterisation and modelling', International Journal for Numerical and Analytical Methods in Geomechanics, vol. 32, no. 11, pp. 1361–1389, 2008.

[33] A. Derkaoui, N. Abou-bekr, and L. Bahi, 'Experimental investigation on the collapse potential of loess soils in Algeria', Bulletin of Engineering Geology and the Environment, vol. 78, no. 6, pp. 4351–4363, 2019.

[34] A. Jotisankasa, W. Mairaing, and D. Tawornwong, 'Collapse behaviour of loess and structured soils: Experimental evidence and interpretation', Engineering Geology, vol. 272, p. 105643, 2020.
[35] Y. Wang, F. Zhang, and H. Liu, 'Shear strength characteristics and collapse mechanism of loess soils', Soils and Foundations, vol. 61, no. 4, pp. 857–870, 2021.

[36] S. Öncü and H. Bilsel, 'Marble powder to stabilise clayey soils in sub-bases for road construction'. 2018.

[37] O. Sivrikaya, E. Aydin, and A. Kaya, 'Effect of marble powder on the geotechnical properties of fine-grained soils', Journal of Materials Research and Technology, vol. 9, no. 5, pp. 11623–11633, 2020, doi: 10.1016/j.jmrt.2020.08.112.

[38] F. Abdelkader, H. Ait Mouheb, and A. Boukhelkhal, 'Stabilization of expansive soils with marble powder', International Research Journal of Engineering and Technology (IRJET), vol. 11, no. 5, pp. 178–185, 2024.

[39] J. E. Jennings and K. Knight, 'A guide to construction on or with materials exhibiting additional settlement due to "collapse" of grain structure', in Proceedings of the 6th Regional Conference for Africa on Soil Mechanics and Foundation Engineering, Durban, 1975, pp. 99–105.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 820-831. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3091

RESEARCH ARTICLE OPEN ACCESS

ENHANCED FACE RECOGNITION SYSTEM INTEGRATING GAN-AUGMENTED CNN FEATURES WITH ICA AND VISION TRANSFORMER

Sangeetha Karunakaran*[1] and A. Akila[2]

[1, 2]Department of Computer Science and Information Technology, School of Computing Sciences, Vels Institute of Science Technology and Advanced Studies (VISTAS), Pallavaram, Chennai, India.

[1]https://orcid.org/0009-0008-3896-8240, [2]https://orcid.org/0000-0001-7744-3473

Email: *sangitanu22@gmail.com, akila.scs@vistas.ac.in

ARTICLE INFO

Article History

Received: December 9, 2025

Revised: January 10, 2026

Accepted: January 15, 2026

Published: February 28, 2026

Keywords:

Face Recognition, Convolutional Neural Network (CNN), Independent Component Analysis (ICA), Vision Transformer (ViT), Generative Adversarial Network (GAN).

ABSTRACT

Facial recognition is a high-technology biometric method that is extensively applied as identity verification, surveillance, and a security system. This paper proposes a hybrid deep learning-based model that can contribute to improved face and iris recognition accuracy and efficiency. This is done by first augmenting the datasets through Generative Adversarial Networks (GANs) that train more synthetic face and iris images to incorporate more and better diversity to the dataset and to reduce overfitting, after which the dataset is fed into a Convolutional Neural Network (CNN) to automatically learn and extract deep spatial features of the augmented images. These extracted features are then narrowed down using Independent Component Analysis (ICA) to select the most important features, removing redundant and irrelevant information. The optimized features are then forwarded to a Vision Transformer (ViT) to be classified by the transformer architecture that takes good consideration of spatial relationships to accurately determine individual face and iris. Performance evaluation metrics of the proposed system include accuracy of 0.93%, precision of 0.938%, recall of 0.930%, and F1-score of 0.9319%, which show that the proposed system has better recognition performance and strength than the conventional face recognition methods.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Recognition Face recognition is one of the most popular research subjects in computer vision and biometrics because it is applied in surveillance, authentication, and human-computer interaction. The field of study has been developing ever since classics of Eigenfaces, Fisher faces, and Local Binary Patterns to modern days deep learning methods that have demonstrated the state of the art [1]. Nevertheless, the variations in poses, changes in the light source, and facial coverages continue to lower the accuracy and reliability of recognition systems in real world situations. Recent surveys indicate that, though deep learning models, notably Convolutional Neural Networks (CNNs), are used in most face recognition studies, the problem of their high computational cost, data scarcity, and overfitting is still a critical issue [2], [3]. To overcome these limitations, there has been a proposal of hybrid approaches which combines both statistical and deep learning models. Other than technical issues, fairness and ethical questions have been discussed in the recent literature as well along with the models that can improve feature selection, generalization, and recognition accuracy when dealing with challenging environments [4]. Recognition accuracy has also been demonstrated to be demographically different, which brings about the concern of biases in automated systems [5].

I.1 OBJECTIVE

- To build a rich face and iris dataset with improvements of data augmentation by Generative Adversarial Networks (GANs) to increase diversity and minimize overfitting.
- To obtain deep facial and iris features with the help of a Convolutional Neural Network (CNN) which acquires the complex spatial and structural patterns of human faces.
- To apply the Independent Component Analysis (ICA) in feature selection, it is necessary to make sure that it is possible to remove redundant and irrelevant features and still have statistically independent components.
- To categorize optimized features with the help of a Vision Transformer (ViT) model that takes into account attention mechanisms to recognize face and iris accurately and efficiently.
- To measure the performance of the system based on the conventional measures of accuracy, precision, recall, and F1-score to confirm the efficiency and applicability of the proposed hybrid model.

I.2 CONTRIBUTION OF THE WORK

- GAN-based augmentation integration will guarantee a bigger, balanced set of data with enhanced variation in face and iris.
- Deep feature extraction (CNN) is an effective method to capture important visual details to enhance the quality of facial representation.
- ICA decreases the number of dimensions of data, increases the efficiency of computation, and increases the independence of features.
- ViT model proposes the use of self-attention mechanism in order to correctly classify across changing lighting and pose.
- The suggested hybrid solution is tested on the basis of major evaluation metrics, which proves to be more accurate and reliable than the classical face and iris recognition models.

I.3 ORGANIZATION OF THE PAPER

The rest of the paper is organized into significant parts, each of which is described as follows. Section II lists the research projects on face and iris Recognition System, completed by various authors. The suggested method's workflow is defined in Section III, and the Results and performance analysis of the face and iris Recognition System are presented in Section IV. Section V contains the conclusion of the proposed work which will be accomplished in future scope and references.

II. RELATED WORK

Face and iris recognition methods to perform authentication and attendance systems have been widely studied in recent times with more focus on performance under practical conditions. In [6], Face ShapeNet (FSN) system was proposed to use in 3D Face Recognition (3D-FR) and in this case, deep shape-aware convolutional neural networks are trained using geometric facial features that are discriminative and directly on 3D scans. The approach enhances resistance to pose and illumination changes by relying on depth and structure data, but its primary limitation is that it needs high quality 3D sensors and more calculations, making it impossible to scale and deploy in low-cost settings in real-time. Continuing on the views of practical deployments, [7] made a Systematic Literature Review (SLR) of Local Binary Pattern Histogram (LBPH) and Convolutional Neural Network (CNN) based face recognition techniques as a student attendance system.

The paper pointed to the fact that CNN-based methods are better in accuracy and ability to adapt to more challenging conditions than LBPH, but those have high training cost, data dependency, and hardware demands whereas LBPH, despite being lightweight, does not perform well in terms of lighting, pose, and occlusion variations. In [8], the implementation of face recognition-based attendance system was done and used with classical computer vision pipelines with Haar Cascade Classifier (HCC) used to detect faces and feature-based recognition systems. The methodology proved to be simple and easy to apply in controlled settings though its weakness is low resistance to change in facial pose, change in lighting and occlusions thus not effective in unconfined or large-scale application.

Discussing the real-world issues that emerge, [9] suggested an effective Masked Face Recognition (MFR) algorithm in the COVID-19 scenario based on the selective analysis of facial parts and feature extraction with periocular focus. Although the method has a significant ability to enhance recognition accuracy of masked faces using less computational load, its drawback is that it has lower discriminative ability where eye-region features are masked or mislead by accessories such as glasses. Adding to masked face recognition, [10] also proposed a Cropping and Attention-Based Network (CABN), which combines Attention Mechanism (AM) and deep learning to pay attention to the unmasked areas of the face. It is an efficient way of enhancing recognition because salient features are dynamically weighted, but it is very sensitive to accurate face alignment and the ability to crop the face and its accuracy is also affected by gross occlusion or incorrect mask identification.

Table 1: Comparison table for related work.

Ref No	Author & (Year)	Data Used	Algorithm Used	Results Achieved (%)
[11]	Ning et al., (2022)	Facial image datasets (public datasets)	Facial feature-based face editing	Identity preservation: ~95%, realistic edits achieved
[12]	Smith & Miller (2022)	Not dataset-focused	Ethical and qualitative analysis	Discussion-based insights;no quantitative results
[13]	Li et al., (2024)	Public face datasets	GPU-accelerated CNN for face recognition	Accuracy: 93–96%, Speed improvement 3–5×
[14]	Damer et al., (2021)	Real and simulated masked face datasets	CNN-based face recognition	Accuracy drops due to masks: 10–15%, Mitigation improved accuracy: 88–90%
[15]	Stevens & Keyes (2021)	Public data & case studies	Sociopolitical and qualitative analysis	Focused on racial bias and ethics; qualitative results

Source: Authors, (2026).

The table 1 is a summary of face recognition studies and other studies conducted recently. It contains work on facial feature-based editing, which does not alter identity, but makes realistic changes with a success rate of about 95%. Ethical and qualitative analyses draw attention to the societal implications, privacy issues and responsible implementation, but do not give any quantitative findings. Convolutional neural networks that are implemented with GPU promote the recognition accuracy to 93-96 percent and accelerates the computing process by 3-5 times. Experiments on masked face recognition reveal that masked recognition accuracy has fallen by 10 -15 percent as a result of masking and masked recognition can be mitigated to approximately 88-90 percent.

III. PROPOSED METHODOLOGY

The face and iris recognition methodology proposed is built on combining both high-level deep learning and statistical methods to ensure high recognition and computational efficiency. It starts by augmentation of the dataset with Generative Adversarial Networks (GANs) to increase the training dataset. GANs create natural synthetic face and iris images, resulting in a multiplicity of data and eliminating overfitting, which improves the capacity of the model to extrapolate between different lighting and pose situations. The augmented dataset is then fed to a Convolutional Neural Network (CNN) which is effective in extracting important facial features like edges, textures, and spatial features. These rich features give a strong depiction of every face and iris in this dataset.

Independent Component Analysis (ICA) is used as a feature selection criterion to optimise the extracted data. The ICA helps to remove the redundant and irrelevant features, and preserve the statistically independent and discriminative elements that would help enhance the model efficiency and decrease the complexity of the computations. This fine-tuned feature set is then subjected to a Vision Transformer (ViT) to make a classification, which employs attention mechanisms to learn spatial relationships and attain an accurate face and iris recognition. Lastly, the system performance is assessed with the help of the key metrics, including accuracy, precision, recall, and F1-score, which proves that the proposed CNN-ICA-ViT hybrid model is superior to the established face and iris recognition algorithms both in reliability and robustness shown in figure 1.

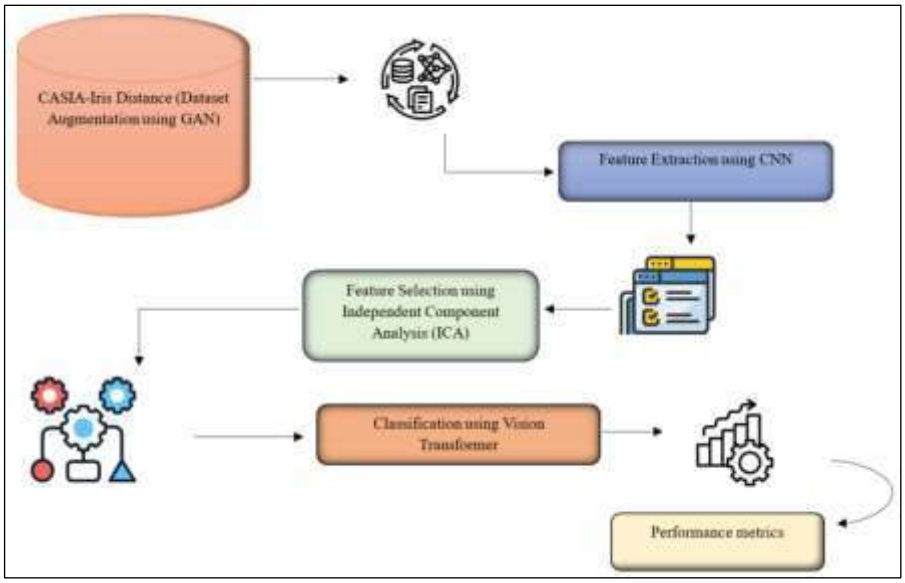


Figure 1: Proposed overall block diagram.
Source: Authors, (2026).

III.1 CASIA-IRIS-DISTANCE DATASET

CASIA-Iris-Distance contains Iris photos captured using our in-house long-range multi-modal biometric image collection. The advanced biometric sensor is able to recognize individuals as far as three meters through active searching of iris, face, or palmprint patterns within the visible field with an intelligent multi-camera imaging system. The LMBS is human-centered and the integration of computer vision, human-computer interface and multi-camera coordination technologies make it highly useful to the existing biometric systems. The picture region of interest is dual-eyed iris, face patterns since the CASIA-Iris-Distance iris images have been captured by a high-resolution camera. Also, in case of multi-modal biometric information fusion, fine-grained features of faces such as skin pattern can be observed.

III.2 DATASET AUGMENTATION USING GENERATIVE ADVERSARIAL NETWORKS (GANS)

Generative Adversarial Networks (GANs) are a high-tech method to augment the existing limited or unbalanced datasets with realistic synthetic samples. A GAN is composed of two networks, one generator, and the other a discriminator that tries to differentiate between fake and real data. Adversarial training enables the generator to learn to generate data that is very realistic and resembles the original data. This will help to improve the performance of models, decrease overfitting, and increase generalization. GAN-based augmentation has been used in image recognition, medical diagnosis, speech analysis, generating large quantities of data is a difficult task and it is a promising method to address data scarcity and imbalance issues.

III.3 FEATURE EXTRACTION USING CONVOLUTIONAL NEURAL NETWORKS (CNNS)

Convolutional Neural Network (CNN) in face and iris recognition is a method of feature extraction, which involves learning and detecting significant facial features in images, including eyes, nose, mouth, and facial structure, automatically. CNNs embrace convolutional, pooling, and fully connected layers to find the local features, edges and textures, dimensionality reduction, and the integration of the features detected on a lower level into the high-level representation respectively. These acquired characteristics are resistant to changes in light, pose and expression, therefore CNNs are very powerful in the ability to differentiate between various face and iris.

$$Z = X * W + b \tag{1}$$

The equation (1) denotes the convolution process of a Convolutional Neural Network (CNN). Convolution operation involves moving the kernel on the input image, multiplying and adding pixels in each receptive field to form the feature map Z. Through this process, the CNN will be able to identify significant local features like edges, corners, and textures, which are significant in a face and iris recognition and feature extraction.

$$f(Z) = \max(0, Z) \tag{2}$$

The equation (2) represents the Rectified Linear Unit (ReLU) activation function used in Convolutional Neural Networks (CNNs). It introduces non-linearity into the model by transforming all negative values in the feature map $Z$to zero while keeping positive values unchanged. This operation helps the network learn complex patterns and relationships within the data. ReLU accelerates the training process by avoiding the vanishing gradient problem, allowing efficient feature extraction. In face and iris recognition, ReLU enhances the model's ability to distinguish subtle facial features such as contours, shapes, and textures, improving overall recognition performance.

$$P(i,j) = \underset{(m,n)\in R(i,j)}{max} z(m,n) \tag{3}$$

The equation (3) represents the max pooling operation in a Convolutional Neural Network (CNN). Here, $Z(m,n)$denotes the input feature map values, and $R(i,j)$defines a small local region (such as 2×2 or 3×3) around the position (i,j). The max pooling function selects the maximum value from each region, producing the pooled feature map $P(i,j)$. This operation reduces the spatial size of the feature maps, decreases computational complexity, and retains the most significant features. In face and iris recognition, it helps preserve dominant facial characteristics while making the model more efficient and robust.

$$P(y = j|x) = \frac{e^{W_j \cdot F + b_j}}{\sum_{k=1}^{K} e^{W_k \cdot F + b_k}} \tag{4}$$

The equation (4) represents the Softmax function, commonly used in the final layer of a Convolutional Neural Network (CNN) for classification tasks like face and iris recognition. Here, Fdenotes the feature vector obtained after convolution and pooling, $W_j$and $b_j$are the weight and bias for class j, and $K$is the total number of classes. The Softmax function converts the network's raw output scores into probabilities that sum to one.

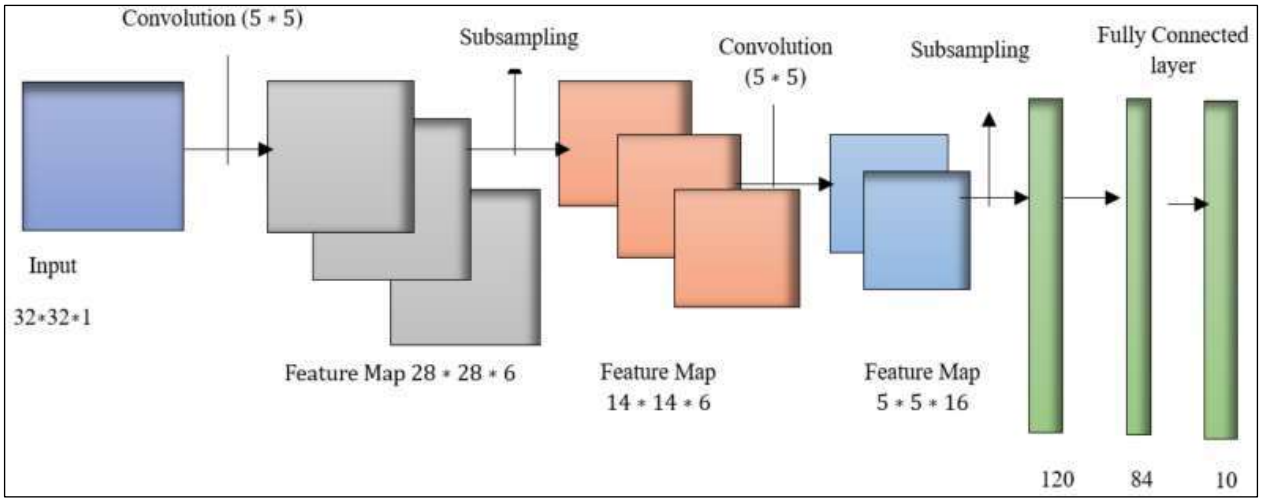


Figure 2: Convolutional Neural Networks Architecture.
Source: Authors, (2026).

The figure 2 demonstrates the structure of a Convolutional Neural Network (CNN) in image processing. The network receives a 32 x 32 x 1 input image which is a grayscale image and initially does a convolution layer using a 5 x 5 kernel which produces a 28 x 28 x 6 channel feature map. This is then subsampled (or pooled) to make the spatial dimensions smaller to extract the main features and obtain spatial invariance. The results of the reduced feature maps are then fed to a second convolution layer with 5x5 kernel which yields another set of feature maps of size 14x14 and 6 channels. This is followed by a subsampling layer that cuts down the size of the feature map, another convolution layer that produces feature maps of 5 X 5 and 16 channels. Lastly, the processed features maps are flattened and run into fully connected layers with 120, 84 and 10 neurons respectively to classify.

III.4 FEATURE SELECTION USING INDEPENDENT COMPONENT ANALYSIS (ICA)

Independent Component Analysis (ICA) is a feature selection method used in datasets to select the most statistically independent features to enhance the efficiency and accuracy of the model. ICA converts the original correlated variables to a new set of independent components, which depict underlying hidden factors on the data. Through these elements, one will be able to eliminate irrelevant and redundant elements, and preserve the most informative elements. This minimizes dimensions, increases interpretability and minimizes computation complexity.

$$X = A.S \tag{5}$$

The ICA equation (5) explains observed data X is formed by mixing hidden independent source signals S through a mixing matrix A. Each column of X represents an observed feature that is actually a combination of several independent sources. Since both A and S are unknown, ICA works by estimating an unmixing matrix W, recovering the original independent components. This process allows the separation of mixed signals into meaningful and independent features, which can then be used for dimensionality reduction and feature selection in deep learning.

$$X_c = X - \mu \tag{6}$$

The equation (6) represents the centering step in Independent Component Analysis (ICA) or Principal Component Analysis (PCA). Here, X is the original dataset, μ\muμ is the mean of each feature, and X_c is the centered data. Subtracting the mean ensures that each feature has a zero mean, which is important for simplifying covariance calculations and improving accuracy in extracting independent or principal components. By removing the mean, the data becomes normalized around the origin, allowing ICA to focus only on variations and dependencies between features rather than being biased by absolute values.

$$X_\omega = VD^{-\frac{1}{2}}V^T X_c \tag{7}$$

The equation (7) represents the whitening (sphering) step in ICA or PCA. Here, X_c is the mean-centered data, V is the matrix of eigenvectors of the covariance matrix, and D is the diagonal matrix of corresponding eigenvalues. The operation $D^{-\frac{1}{2}}$ scales the data so that each transformed feature has unit variance. Multiplying with V and V^Tensures that the features are uncorrelated. Thus, whitening transforms correlated variables into a new set of orthogonal features with zero mean and unit variance, preparing the data for extracting independent components in ICA.

III.5 FACE AND IRIS RECOGNITION USING VISION TRANSFORMER-BASED CLASSIFICATION

Vision Transformer (ViT) is a form of face and iris recognition which can be used as an alternative to conventional convolutional neural networks (CNNs). ViT model splits a facial image into small patches and treats them as sequences as the words in natural language processing. A patch is first turned into an embedding and self-attention mechanisms are used to learn global relations between face and iris features. The method allows the model to acquire finer details and elaborate facial patterns effectively.

$$x_p = \left[x_p^1, x_p^2 \ldots . x_p^N\right], x_p^i \in \mathbb{R}^{P^2.C} \tag{8}$$

The equation (8) represents the image patch embedding process in the Vision Transformer. Here, the input image is divided into N non-overlapping patches of size $P \times P$ with Ccolor channels. Each patch is flattened into a one-dimensional vector, allowing the model to treat the image as a sequence of patch embeddings for further processing.

$$z_0 = \left[x_{class}; x_p^1 E; x_p^2 E; \ldots x_p^N E\right] + E_{pos} \tag{9}$$

The equation (9) represents the initial input embedding in the Vision Transformer model. Here, each image patch $x_p^i$is linearly projected using an embedding matrix E. A special classification token $x_{class}$is added at the beginning of the sequence, and positional embedding $E_{pos}$is included to preserve spatial information, enabling the model to learn the order and relative position of patches effectively.

$$Z' = MSA\big(LN(Z)\big) + Z \tag{10}$$

The equation (10) represents the Multi-Head Self-Attention (MSA) operation within the Vision Transformer encoder. Here, LN(Z)denotes layer normalization applied to the input embeddings Z, which helps stabilize and speed up training. The MSA mechanism then captures relationships between all image patches by attending to different parts of the sequence simultaneously.

$$y = \text{Softmax}(W_o . z_{class}^{(L)} + b_o) \tag{11}$$

The equation (11) represents the final classification step in the Vision Transformer model. Here, $z_{class}^{(L)}$is the output of the classification token from the last Transformer encoder layer, which contains the global image representation. $W_o$and $b_o$are the weight matrix and bias term of the output layer. The Softmax function then converts the computed scores into probability values, allowing the model to classify the input face and iris image into the correct identity category.

IV. RESULTS AND DISCUSSION

A hybrid face and iris recognition model was tested with the use of standard performance measures. The outcomes showed that there were a high improvement of the recognition accuracy and the overall efficiency. The system had a balanced performance with the values of accuracy, precision, recall and F1-score being almost equal. These results confirm the usefulness of the combined GAN, CNN, ICA, and ViT strategy in comparison to the traditional approaches.



Figure 3: GAN Data Augmentation.
Source: Authors, (2026).

A Generative Adversarial Network (GAN) consists of two models, which are a Generator and a Discriminator shown in figure 3. The generator produces simulated data by using deep neural networks to make synthetic data samples out of random noise and the discriminator to classify inputs as real or fake is a convolutional-based network. The adversarial training and minimax optimization help the networks to improve each other, as the generator network is trained to generate natural images and the discriminator is trained to be more accurate at identifying these images. This makes it possible to produce a variety of high-quality synthetic images, which can be used to scale the small datasets.

Data augmentation with GANs enhances the generalization of the model, its strength, and accuracy of classification in medical and image analysis problems. The figure (4) is a representation of the signal or score of the Independent Component Analysis (ICA) Component 1. The x-axis is a sample index, and the y-axis is a component amplitude of approximately +/-3 to +/-4. These fluctuations are indicators of variations of the independent source signal extracted in the mixed dataset. ICA separates multivariate data to find statistically independent parts that show concealed patterns and noise-free signals. This is essential in the extraction of features, separation of signals and removal of artifacts in EEG, fMRI, and MRI studies, and it improves data-interpretation and classification accuracy in the downstream analysis model.

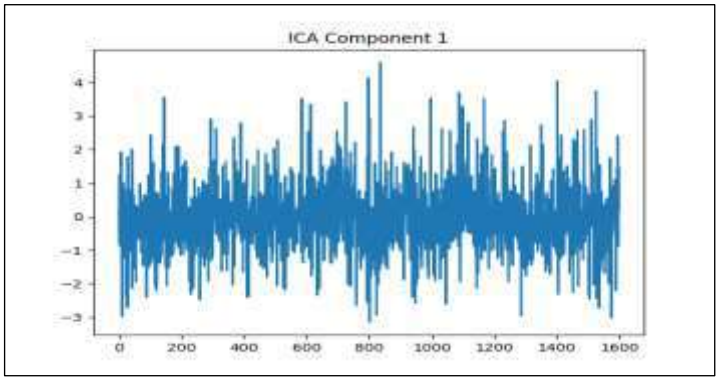


Figure 4: ICA Component 1.
Source: Authors, (2026).

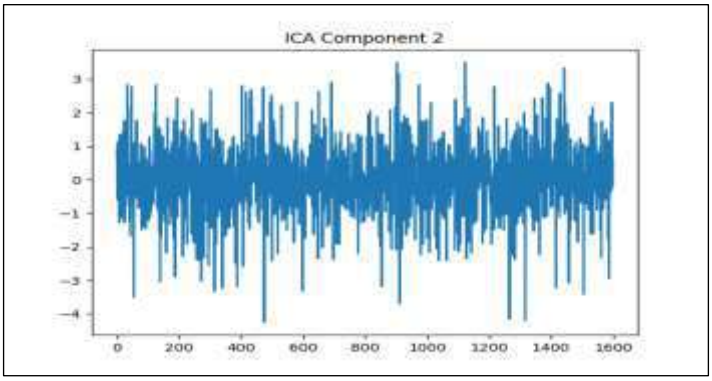


Figure 5: ICA Component 2.
Source: Authors, (2026).

The second element that was extracted on the dataset features by the Independent Component Analysis (ICA) is the figure 5. The x-axis is representing each sample in the dataset and the y-axis represents the value of the component. The values vary between -4.5 and 3.5 which shows the variation and dispersion of this independent source on all the samples. Peaks and troughs retrieve different patterns that exist in the data, showing variation that are statistically independent of other components. This kind of representation can be used to reduce dimensionality and extract features with the subsequent classification as it isolates non-redundant information.

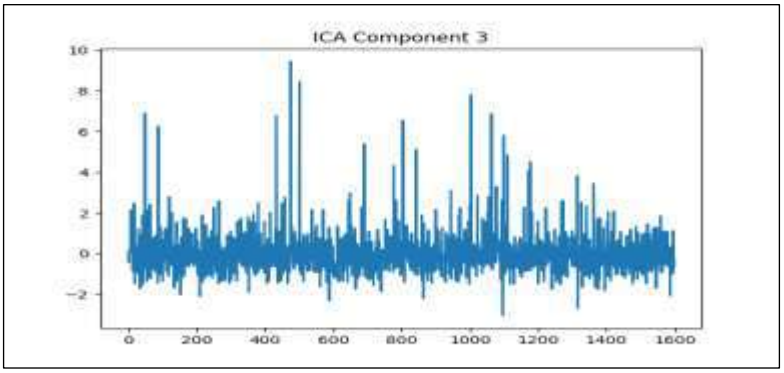


Figure 6: ICA Component 3.
Source: Authors, (2026).

The figure 6 depicts ICA Component 3 that is among the independent components generated by Independent Component Analysis (ICA). The sample index varies approximately between 0 and 1600 as demonstrated by the x-axis and the component amplitude varies between around -3 and 10 as illustrated by the y-axis. The signal swings around zero and we can see the multiple sharp peaks with a value exceeding 8, which means that there is a strong independent activity at these points. The majority of the data values fall within the range of -2 and 3 indicating that the component is mostly composed of noise with low amplitude with occasional large spikes that may be indicative of big independent signals that are not mixed.

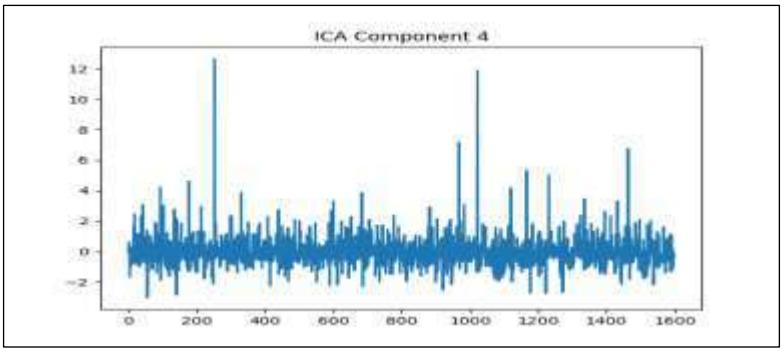


Figure 7: ICA Component 4.
Source: Authors, (2026).

The figure 7 ICA Component 4 which is one of the independent signals obtained through Independent Component Analysis (ICA). The x-axis will go between 0 and 1600 which is the index of the sample, the y-axis will go between approximately -3 and 13 which is the amplitude of the component. The signal values mostly lie in the range of -2 to 3, indicating that they are stable and have a low-level background activity. Nevertheless, it has some major spikes that possess a value above 10, especially at sample indices 250 and 1000, which show strong independent events or artifacts in the signal. These maxima can be associated with different underlying sources which have been isolated by the ICA process.

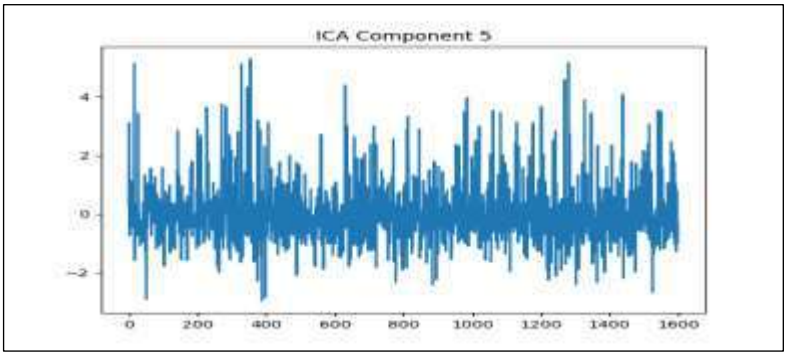


Figure 8: ICA Component 5.
Source: Authors, (2026).

The figure 8 depiction of the ICA Component 5 that is one of the independent signals that are isolated using Independent Component Analysis (ICA). The x-axis goes between 0 and 1600 indicating the index of the samples and the y-axis is approximately between -3 to 5 indicating the gain of the signal. Fluctuations around the zero line have a very high density which means that the waveform varies randomly, such that there are occasional sharp peaks which extend up to +5 and dips close to -2.5. The maximum portion of data is found between -1 and 3 indicating medium signal intensity. Such fluctuations can be an independent source having a stable background activity and moderate excursions of higher signal elements.

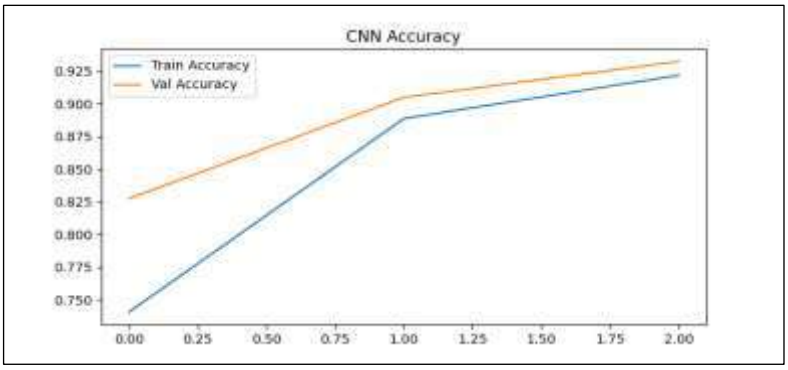


Figure 9: CNN Accuracy Explanation.
Source: Authors, (2026).

The figure 9 CNN accuracy graph, the model has been trained and validated 3 times. The accuracy of the training begins at approximately 0.74 and it gradually increases to approximately 0.92, which means that the model is learning efficiently. Validation accuracy starts with a high value of about 0.83 and increases up to approximately 0.93 indicating good generalization on unknown data.

Training and validation curves are very close indicating that the model is not overfitting. The steady improvement in epochs confirms the CNN is indeed reducing the error and enhancing feature extraction with every training round resulting in good classification accuracy and ideal predictive behavior.

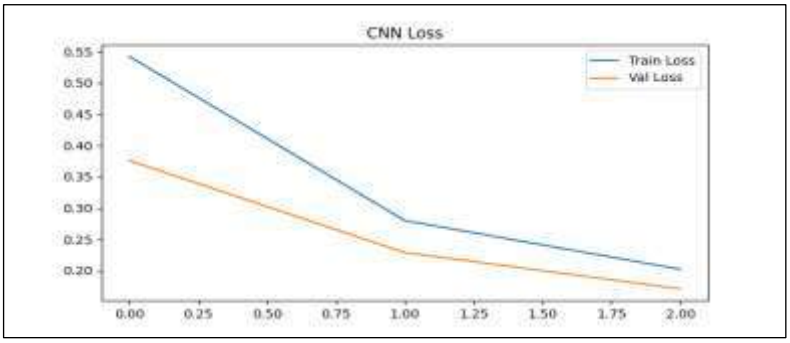


Figure 10: CNN Loss Explanation.
Source: Authors, (2026).

The CNN loss curve shows the curve of training and validation losses of the model in relation to the epochs and it shows an improvement. There would be an initial high loss of 0.55 and then a decrease to approximately 0.20 on the third epoch which is a sign of successful optimization. On the same note, validation loss reduces by about 0.38 to 0.16, which proves that the model is acquiring general patterns. This declining pattern of both the losses is an indication that the CNN is on the proper course. The reduced difference between training and validation losses indicates low levels of overfitting and stable learning. In general, the error minimization proves that the CNN is capable of reducing the prediction errors and attaining more accurate models shown in figure 10.

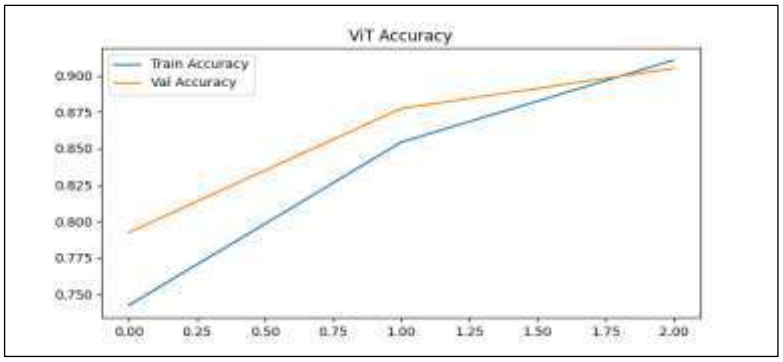


Figure 11: ViT Accuracy Explanation.
Source: Authors, (2026).

The ViT (Vision Transformer) accuracy graph indicates the training and validation accuracy after three epochs and the accuracy of the model is gradually improving. The training accuracy starts with a value of about 0.74 and then it starts to increase to about 0.91 and the validation accuracy starts with about 0.79 then it slightly rises to about 0.90. This overall positive result in all epochs shows that the ViT model is effective in learning image representations with the help of attention mechanisms. There is close relation between training and validation accuracy curves implying that there is less overfitting of the model that is applied to unseen data. The high results emphasize the ability of the ViT to express the characteristics of complex images and obtain a high classification rate shown in figure 11.

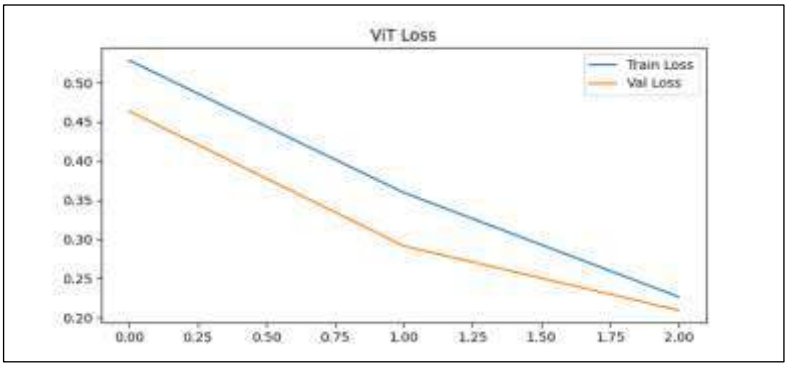


Figure 12: ViT Loss Explanation.
Source: Authors, (2026).

The ViT loss graph indicates improvements in the performance of the model with a reduction in training and validation loss with an increase in the number of epochs. The training loss commences at about 0.50 and gradually decreases to about 0.20 whereas the validation loss commences at about 0.45 and also decreases to about 0.18. This gradual decrease shows that the Vision Transformer is an effective way to reduce the error in predictions with the help of its attention-based learning mechanism. The two loss curves are close which indicates that there is a strong generalization and low overfitting. All in all, this decrease in the percentages of the losses shows that the ViT model is effectively tuning its parameters and attaining more stable and more accurate learning shown in figure 12.

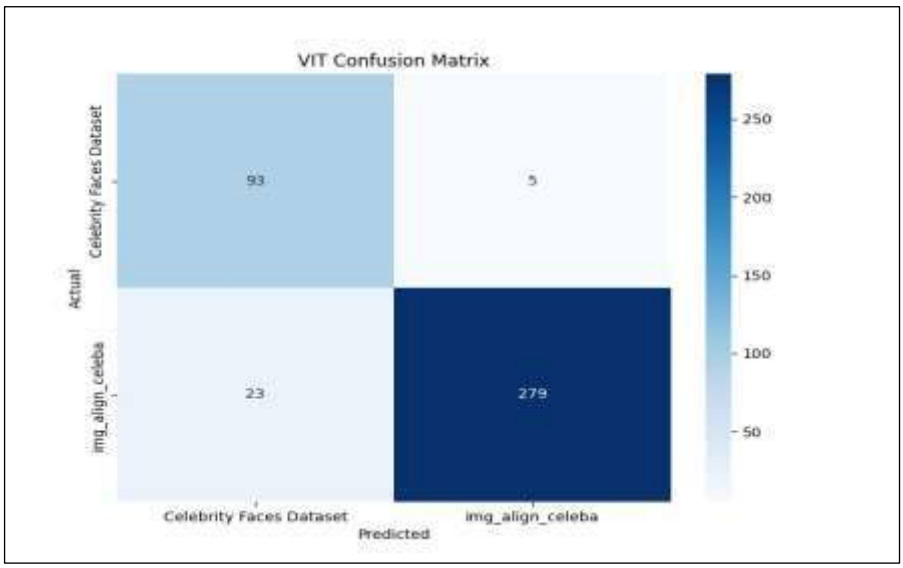


Figure 13: VIT Confusion Matrix.
Source: Authors, (2026).

The figure 13 presented confusion matrix is the assessment of a Vision Transformer (ViT) model that is used in the dataset: "CASIA-IRIS Dataset. The matrix shows that the model was correct in the prediction of 93 examples of the CASIA-IRIS Dataset. It falsely identified 5 images of CASIA-IRIS dataset.

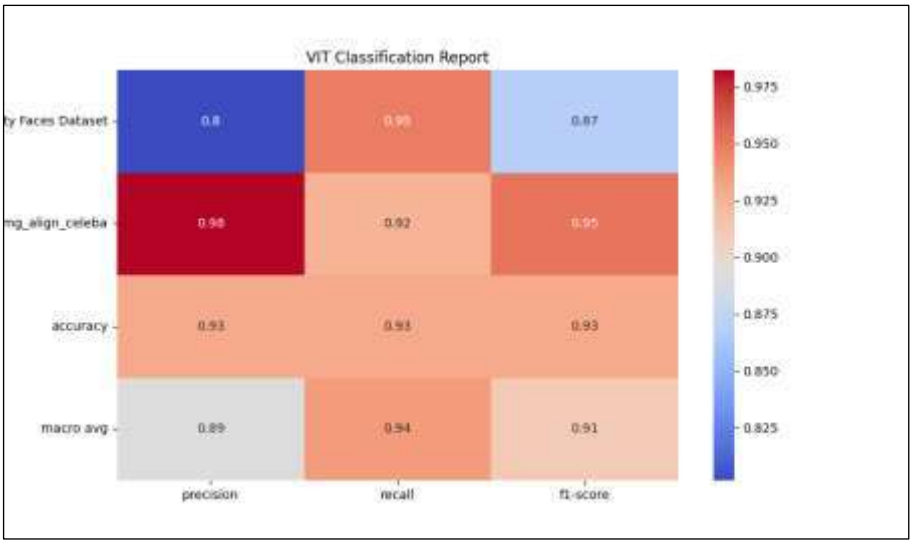


Figure 14: VIT Classification Report.
Source: Authors, (2026).

The figure 14 is a classification report heatmap of a Vision Transformer (ViT) model that summarizes its performance on two classes, namely, “CASIA-IRIS Dataset”. Cutting edge performance measures are precision, recall and F1-score. The model performed well with high precision (0.98) and F1-score (0.95) on CASIA-IRIS dataset. But in the case of the Celebrity Faces Dataset, the precision was small with 0.80 whereas the recall was high with 0.95. The accuracy of the entire model is 0.93. The mean of the two classes on the macro level has balanced recall (0.94) and F1-score (0.91), but a little less precision (0.89) of the CASIA-IRIS Dataset compared to the other class.

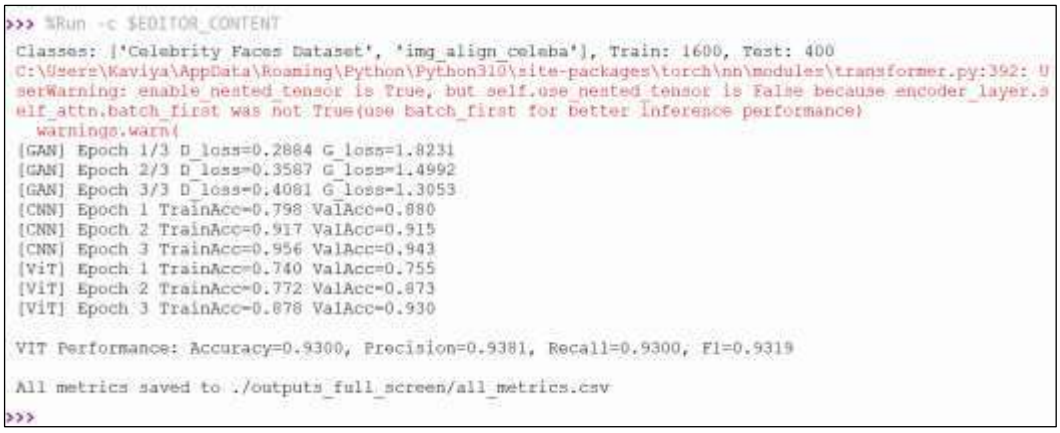


Figure 15: VIT Performance metrics.
Source: Authors, (2026).

The figure 15 output provides the training and evaluation of three models: GAN, CNN, and ViT on two classes, namely, Celebrity Face and img_align_celeba. ViT model reached a high accuracy of 93.00, precision of 93.81, recall of 93.00 and F1-score of 93.19, which means that its results are balanced and excellent. Accuracy is the overall measure of correctness whereas the measure of precision is the moment when the model does not produce false positives. Recall reflects its ability to capture real positives and F1 strikes a balance between the two. The training accuracy of the ViT rose to 87.8 (and the validation accuracy rose to 93.2) in three epochs with a declining loss and steady learning improvement.

V. CONCLUSION

In this paper, a hybrid facial and iris recognition system using deep-learning was introduced with the combination of Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNN), Independent Component Analysis (ICA), and Vision Transformers (ViT). GANs were useful in data augmentation to enhance diversity in the dataset and minimize overfitting. CNNs obtained deep spatial features, whereas ICA optimized features choice eliminating redundancy. Lastly, ViT was able to attain correct classification through modeling spatial dependencies. The system exhibited good performance scored high on accuracy (0.93%), precision (0.938%), recall (0.930%) and F1-score (0.9319%) which is better than the traditional recognition techniques.

The future studies might deal with the incorporation of edge computing-based real-time facial recognition that has a faster processing rate and lower latency. Also, further improvements of model generalization can be done by using more sophisticated approaches to data augmentation and training with adversarial loss. It can be made more applicable by extending the model to work with occlusions, aging effects and cross-domain datasets (e.g. thermal or infrared images). Lastly, by deploying privacy-preserving solutions, e.g., federated learning, one might be able to use it safely and in a responsible way in sensitive domains such as healthcare and law enforcement.

VI.AUTHOR'S CONTRIBUTION

Conceptualization: Sangeetha Karunakaran and Dr.A. Akila.
Methodology: Sangeetha Karunakaran and Dr.A. Akila..
Investigation: Sangeetha Karunakaran and Dr.A. Akila..
Discussion of results: Sangeetha Karunakaran and Dr.A. Akila.
Writing – Original Draft: Sangeetha Karunakaran and Dr.A. Akila.
Writing – Review and Editing: Sangeetha Karunakaran and Dr.A. Akila.
Resources: Sangeetha Karunakaran and Dr.A. Akila.
Supervision: Sangeetha Karunakaran and Dr.A. Akila.
Approval of the final text: Sangeetha Karunakaran and Dr.A. Akila.

VII. REFERENCES

[1] Ali, Waqar, Wenhong Tian, Salah Ud Din, Desire Iradukunda, and Abdullah Aman Khan. "Classical and modern face recognition approaches: a complete review." Multimedia tools and applications 80, no. 3, 2021.

[2] Gururaj, H. L., B. C. Soundarya, S. Priya, J. Shreyas, and Francesco Flammini. "A comprehensive review of face recognition techniques, trends and challenges." IEEE Access,2024.

[3] Wang, Zhongyuan, Baojin Huang, Guangcheng Wang, Peng Yi, and Kui Jiang. "Masked face recognition dataset and application." IEEE Transactions on Biometrics, Behavior, and Identity Science 5, no. 2, 2023.

[4] Gururaj, H. L., B. C. Soundarya, S. Priya, J. Shreyas, and Francesco Flammini. "A comprehensive review of face recognition techniques, trends and challenges." IEEE Access, 2024.

[5] Terhorst, Philipp, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales Moreno, Julian Fierrez, and Arjan Kuijper. "A comprehensive study on face recognition biases beyond demographics." IEEE Transactions on Technology and Society 3, no. 1, 2021.

[6] Jabberi, Marwa, Ali Wali, Bilel Neji, Taha Beyrouthy, and Adel M. Alimi. "Face shapenets for 3d face recognition." IEEE Access 11, 2023.

[7] Budiman, Andre, Ricky Aryatama Yaputera, Said Achmad, and Aditya Kurniawan. "Student attendance with face recognition (LBPH or CNN): Systematic literature review." Procedia Computer Science 216, 2023.

[8] Ramdhon, Andri Nugraha, and Fadly Febriya. "Penerapan Face Recognition Pada Sistem Presensi." Journal of Applied Computer Science and Technology 2, no. 1 (2021): 12-17.

[9] Hariri, Walid. "Efficient masked face recognition method during the covid-19 pandemic." Signal, image and video processing 16, no. 3, 2022.

[10] Li, Yande, Kun Guo, Yonggang Lu, and Li Liu. "Cropping and attention-based approach for masked face recognition." Applied Intelligence 51, no. 5, 2021.

[11] Ning, Xin, Shaohui Xu, Fangzhe Nan, Qingliang Zeng, Chen Wang, Weiwei Cai, Weijun Li, and Yizhang Jiang. "Face editing based on facial recognition features." IEEE Transactions on Cognitive and Developmental Systems 15, no. 2 (2022): 774-783.

[12] Smith, Marcus, and Seumas Miller. "The ethical application of biometric facial recognition technology." Ai & Society 37, no. 1 (2022): 167-175.

[13] Li, Huixiang, Ang Li, Yuning Liu, Yiyu Lin, and Yadong Shi. "AI face recognition and processing technology based on GPU computing." Journal of Theory and Practice of Engineering Science 4, no. 05,2024.

[14] Damer, Naser, Fadi Boutros, Marius Süßmilch, Florian Kirchbuchner, and Arjan Kuijper. "Extended evaluation of the effect of real and simulated masks on face recognition performance." Iet Biometrics 10, no. 5 (2021): 548-561.

[15] Stevens, Nikki, and Os Keyes. "Seeing infrastructure: Race, facial recognition and the politics of data." Cultural Studies 35,2021.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 832-845. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3092

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

RISK-BASED PERFORMANCE OPTIMIZATION OF MOZZARELLA CHEESE PROCESSING USING EXPERT JUDGEMENT – PREFERENCE SELECTION INDEX (PSI) MODEL

Vaishali Talkit*[1], Anand Bewoor[2] and Amitkumar Shelar[3]

[1]Department of Mechanical Engineering, Zeal College of Engineering, Savitribai Phule University, Pune 411041, Maharashtra, India.
[2]Department of Mechanical Engineering, Cummins College of Engineering, Pune 411044, Maharashtra, India
[3]Department of Automobile Engineering, SVPM's ITE Malegaon Bk., Pune 413115, Maharashtra, India

[1]http://orcid.org/0000-0003-1465-5940, [2]http://orcid.org/0000-0001-9329-3807, [3]http://orcid.org/0000-0002-0006-8094

Email: *vaishali.talkit15@gmail.com, dranandbewoor@gmail.com, amitkumarshelar@gmail.com

ARTICLE INFO

Article History
Received: December 9, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Failure Mode Effect and Criticality Analysis (FMECA),
Risk Priority Number (RPN),
Expert Judgement,
PSI model,
Risk Assessment,
Food Quality and Safety

ABSTRACT

The failures of machines in the food processing plants tend to cause lower productivity, quality decline, and increased operation costs as a result of the unplanned downtime and loss of products. The production of mozzarella cheese is extremely sensitive to the reliability of equipment since its production is a thermal-mechanical process that needs to be controlled on a regular basis. This paper introduces a combined Expert Judgement-Preference Selection Index (EJ-PSI) model to consider and prioritize risks related to machine failures in the Mozzarella cheese processing. The relevance of decision criteria to the weighted relevance of the decision criteria was assigned by quantifying and normalizing expert evaluations of ten industry experts on the basis of experience, qualification, supervisory exposure, and audit involvement. The model considers three key systems, eleven key components, and twenty-one key failure modes, thus producing ranked risk profiles to prioritize maintenance. The findings show that the Moulder, Pre-Cooking & Cooking Stretching Machine and Stretching Machine are the most important systems that determine the throughput and the severity of defects. The EJ-PSI model facilitates the predictive maintenance planning process by determining the components that need to be addressed in time to prevent possible failures. The given model proves to be highly applicable in the context of complex production settings as it allows making decisions based on risks and minimizes variability in operations.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The well structured and managed productive environment is vital to the production of high quality cheese that is in line with customer expectations and regulatory standards. The food industries are crucial in the management of the requirement of nourishment of the human beings. Food production, food processing, food packaging, food distribution and consumption are all included in the food industries. Government [1] has ignored the issue of increased population and its effect on food producing industries to a large extent. Food processing industry has a high rate of growth because of the increasing demand of high quality products. The standards such as Good Manufacturing practises (GMP), ISO 9000/22000, Hazard Analysis and Critical Control Points (HACCP), and Good Hygienic practices (GHP) are important to maintain the quality of the products and remain globally competitive. These practices are conducive to the health of consumers, increase the product acceptance, and correspond to the international food safety and hygiene standards. The ultimate goal is to drive the food industry to greater and more confident quality and safety measures. The contemporary food plants are operated on a round-the-clock basis, and even the failure of a single equipment may interfere with the production or the quality of the products.

This renders the planned and well-managed maintenance a necessity in order to regulate the downtime, ensure the quality remains constant, and to ensure the necessary output levels are maintained. Despite the popularity of FMEA in the field of mechanical and manufacturing industries to learn potential failures of machines and systems, its application in food processing remains rather limited. As the standards of safer and more high-quality food have increased, the application of FMEA to the sphere has become more significant. When correctly applied, it will assist in identifying the weak areas in the production line, aid in timely maintenance measures, and aid in efficient production as well as safer food handling. Lack of proper maintenance practices may have direct impacts on the quality and safety of food products in manufacturing plants. Research indicates that poor maintenance can severely reduce equipment efficiency; in a study case, a food company was operating at 55.30 percent efficiency, which is far below the accepted international standards [2]. Such a decline in efficiency increases the possibility of safety errors and quality flaws.

To prevent such problems, the companies should reinforce their maintenance strategy in such a way that machines do not fail and production is not disrupted by unnecessary causes [3]. The use of Failure Mode and Effect Analysis (FMEA) in food processing has gained more significance in enhancing the quality of products, safety, and facilitating operations without any hitches. FMEA is used in combination with other methods like Fault Tree Analysis and HACCP to identify the weak areas of equipments and processes in a systematic manner [4], [5]. By [6] emphasized on increased maintenance in food production and that the failure of equipment may lead to spoilage of the products and scraps in large volumes during line stops. Their research involved a review of failure and repair data of four pizza production lines, which involved descriptive statistics and compared performance among the lines.The findings revealed that time-to-failure and time-to-repair were both lognormally distributed, and the time-to-repair exceeding approximately 45 minutes was an indication of significant operational issues.

Failure Mode Effect Criticality Analysis (FMECA) is an analysis method that integrates the principles of FMEA with a criticality analysis that quantifies the impact of each failure mode on the performance of the system and safety of the product. The same strategy was used when a study on coconut oil processing was conducted and the criticality priority number was significantly improved by about 48.33 percent with the application of FMECA demonstrating the effectiveness of such method in refining maintenance and safety practices [7]. Multi-Criteria Decision Making (MCDM) is a field that is concerned with the assessment of multiple competing factors to assist decision-makers in selecting the most appropriate one in complicated cases.These techniques are used to make sure that the decisions are balanced, transparent and consistent by introducing various criteria into a single framework. In one of the applications, the Analytic Hierarchy Process (AHP) was applied to rank eight possible business partners based on financial and non-financial aspects [8].

AHP is still used to solve multi-criteria problems, as it assists the user to de-compose complex decisions and select the most sensible one, despite some weaknesses indicated [9]. Top Sigma is another method, which uses TOPSIS to rank options using an ideal and a worst-case reference. It is valued because it is easy and user-friendly though it may have difficulties with the weight assignment, disregarding relations between attributes, and can cause rank-reversal effects in certain cases [10], [11]. It may also be utilized in a group or in case of interval based data. Another popular technique is the Weighted Sum Model (WSM) [12]. The simplicity of its structure and speed of calculations make it appropriate when decisions are to be made with multiple criteria with various degrees of significance.The traditional FMEA has few disadvantages. The biggest problem is that it considers all the risk factors as having the same level of significance, which may misrepresent the significance of a failure mode.

Also, various combinations of severity, occurrence and detection may yield the same RPN value and it is more difficult to identify the issues that really require attention. FMEA is only applied on a rudimentary basis in the food industry. Small and medium food processors have many untrained employees, are not aware of structured risk-assessment tools, and do not have the time to analyze them in detail. The industry has its own complications as well like fluctuation of raw materials, strict hygiene regulations and high regulatory requirements that render the regular use of FMEA more difficult than in mechanical industries. In order to address these shortcomings, scholars have resorted to Multi-Criteria Decision-Making (MCDM) techniques, which are more flexible and more precise in ranking the risks [13]. One of the studies investigated the PSI approach to machining-related decisions, including the selection of materials or the cutting fluids, and compared it to other popular MCDM methods [14]. By adding the Expert Judgment (ExJ) component to the PSI model, it is possible to minimize the individual bias by combining the views of several experts and make the evaluation more balanced and representative.

Another issue that PSI approach circumvents is the issue of identical RPN value because normalized scores and weighted criteria are used to obtain clearer and more discriminative rankings. Reliability, Availability, and Maintainability (RAM) analysis is also essential in food production since it allows to reduce downtime, enhance the quality of products, and assist in safe and efficient operation [15], [16]. The current paper suggests a superior model that integrates the expert judgment and PSI method in an FMECA model. This is a combined methodology combining qualitative expertise and quantitative analysis to determine critical failure modes, establish realistic priorities and propose improvements. When it comes to a mozzarella cheese plant, the use of such MCDM-based methods will allow allocating resources more effectively, making more informed decisions, and enhancing the efficiency and sustainability. In this research work, Failure Mode Effect and Criticality Analysis (FMECA) model is a feasible method of analyzing risks and determining potential failure modes in systems, subsystems, or single components.

FMEA, FMECA, and MCDM techniques with an Expert-Judgment-based PSI method are the tools that can be used in a mozzarella cheese plant and each of them has its own benefits in risk assessment and decision support. FMEA can be used to identify possible failures in the production line and prioritize them according to their severity, frequency and detectability. It assists during the initial phase of risk assessment, but its findings may be affected by subjective scoring as well as its simplified assumptions. FMECA expands on FMEA with a criticality assessment, which enables a better assessment of the likelihood of each failure occurring and the intensity with which it can impact product quality or equipment performance. This renders it useful in detecting significant issues, including contamination or mechanical issues, but it may need considerable data and analysis. In the meantime, the MCDM tools like AHP, TOPSIS and VIKOR, with the assistance of the PSI method, offer a more comprehensive and flexible decision-making framework. The inclusion of expert opinions in the PSI process assists in handling multi-factor complex decisions, be it in the selection of suppliers, production planning, or sustainability enhancement.

This combination enables the consideration of both numerical and expert data, leading to more balanced decisions and more in line with the needs of operations. A combination of these methods will result in a more balanced approach to risk management and better performance of the plant. The Expert-Judgment PSI technique also solves a number of shortcomings of the traditional FMEA since it does not produce RPN-related discrepancies, lessens subjectivity, and provides more precise and dependable rankings

II. METHODOLOGY

Combining Expert Judgement (ExJ) with the Preference Selection Index (PSI) approach offers a powerful and systematic framework of ranking failure modes in mozzarella cheese production systems. The methodology is based on the domain knowledge of highly experienced staff members and at the same time a quantitative decision-evaluation framework is utilized. During the initial step, ExJ absorbs the experience of risk perception of the experts due to the operational exposure, machine behaviour, past failures and corrective measures in the industrial practice. These professional judgments are then transformed into decision matrices with regard to chosen criteria like Severity (S), Occurrence (O) and Detection (D), which minimizes the ambiguity in qualitative judgments. The PSI approach is then used to convert normalized expert data into objective risk-based ranking.

In contrast to traditional risk assessment methods that depend on the subjective scoring only, PSI removes the problem of inconsistency by calculating the variation of preference, the relative proximity to optimal performance, and the ranking of the components. This strategy reduces personal bias, prioritization of failures can be reproduced, and weight distribution is consistent. This integrated ExJ-PSI framework can be used in food processing systems where variation in quality and safety violations are caused by deviations in operations, like mozzarella cheese lines where the subsystems and components that are more likely to fail can be identified. This kind of prioritization leads to predictive maintenance scheduling, less downtime and better production reliability. Figure 1 shows the sequential methodological steps followed in this study that are data collection, expert scoring, normalization, PSI computation, and final ranking. The sequence of steps used in this analysis is illustrated in Figure 1.

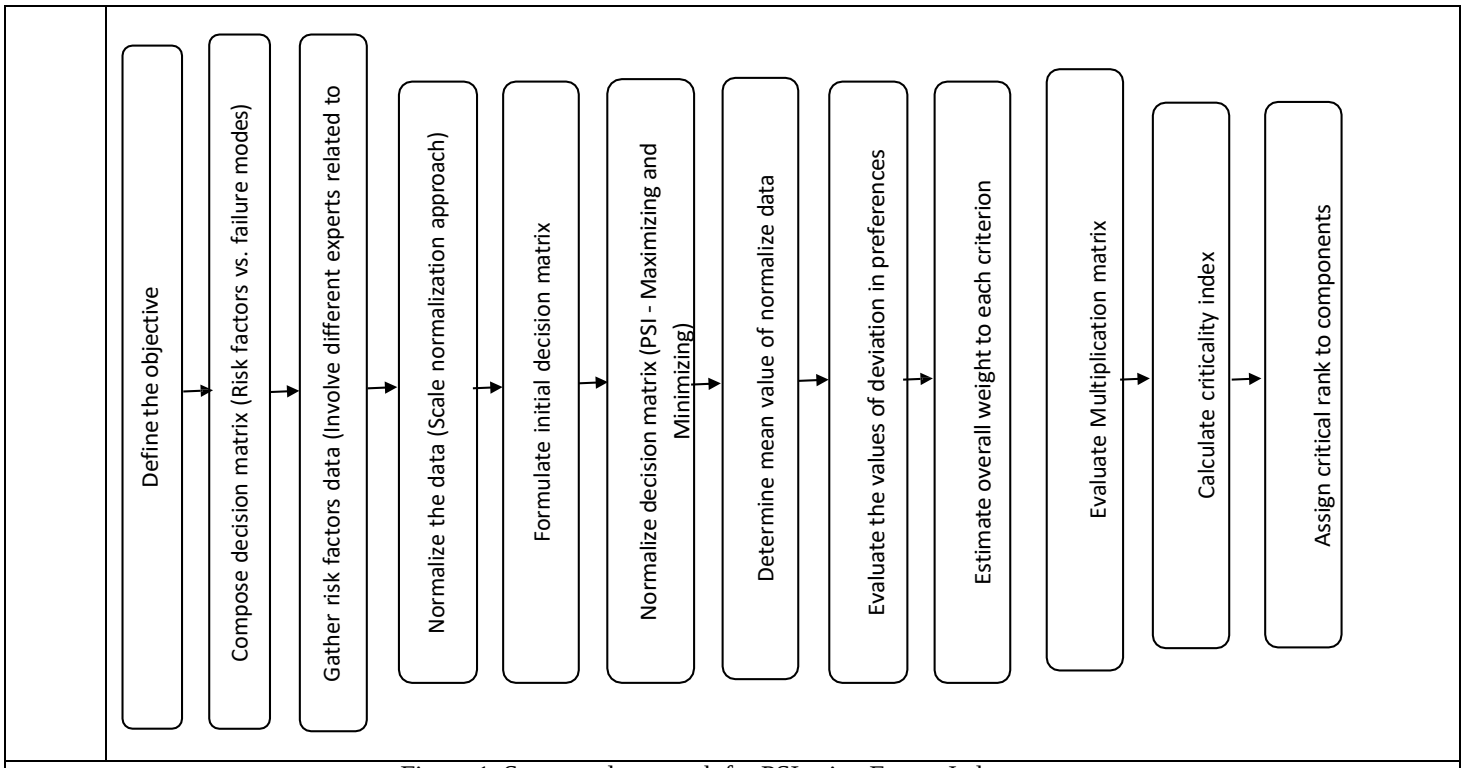


Figure 1: Structured approach for PSI using Expert Judgments.

Source: Authors, (2026).

III. CASE STUDY OF SUBSYSTEMS/COMPONENTS IN MOZZARELLA CHEESE PLANT

The case study is devoted to the assessment of key subsystems and key machine parts of a mozzarella cheese processing unit of an industry with a special emphasis on the determination of failure modes which directly affect productivity, reliability, and the quality of products. The process of mozzarella production has several consecutive and mutually dependent stages, including the creation of curd, pre-cooking, stretching, molding, cooling, brining and packaging. Each phase is assisted by specific machines and auxiliary systems whose operation leads to the overall performance of processing, energy consumption, and adherence to standards of consistency and hygiene.

Criticality analysis In order to examine the criticality, a combined Expert Judgement-Preference Selection Index (EXJ-PSI) framework was introduced to conduct systematic evaluation of the subsystem reliability. Technical operators, maintenance engineers and supervisory personnel who were well versed with the operational bottlenecks, historical breakdown trends and processes which were sensitive to safety were all inputted as experts. The obtained information was mapped to component-based failure modes, severity indices, detection patterns, and downtime impacts, which would be realistic to failure representation. Using EXJ-PSI, the research prioritizes machine-level risks that cause process interruptions, performance losses and non-uniformity of products.

The process that was followed in identifying subsystems, assessment, grading, standardization, and ranking is outlined in the following section. This systematic case- based study is the foundation of the maintenance planning, equipment tracking, and failure prevention in the processing of mozzarella cheese..Write in detail the research project, including background and limitations. The selection of materials and methods, procedures and equipment must be justified so that the work can be reproduced. Modifications or new methods must be described in detail. You must clearly define the universe and specify how the sample was selected and why it is representative. Data processing represents the practical development of a theoretical basis, deriving the model equations to program the calculation algorithm, according to the need. In materials, they include the technical specifications and the quantities, the origin and, if necessary, the method for its elaboration.

III.1 STEPS FOLLOWED FOR INTERGRATED EXJ-PSI (EXPERT JUDGEMENT – PREFERENCE SELECTION INDEX)

The three decision criteria were identified based on the severity of the failure mode (S), occurrence (O) and detection (D) which were summarized in Table 1, Table 2 and Table 3 respectively. The criteria reflect the extent of impact, the risk of failure recurrence and the chance of early detection respectively. The suggested integrated ExJ-PSI model proves to be very applicable in the analysis of complex equipment structures that have many subsystems, many components, and interacting failure modes [17], [18]. The methodological approach taken in this study is based on the methodological principles of the previous research by [19] and it is given as follows:

1. Decision Matrix Formulation - Determine the failure modes (FM) and evaluation criteria, i.e. Severity (S), Occurrence (O) and Detection (D), which were used to prioritize risks.
2. Expert Input Compilation - Gather information of experts and specialists in the domain and industry with similar machine setups and operational history, so that realistic depiction of past failures and maintenance patterns can be obtained.

Table 1: FMECA scale employed to identify severity (S).

Severity Level	Category	Description	Rank
1	Negligible	No noticeable effect on cheese quality, safety, or production.	1
2	Low	Minimal effect on cheese production, easily corrected.	2
3		Slight impact on production efficiency or cheese quality, but no major delays or rework needed.	3
4	Moderate	Reduced efficiency or minor impact on cheese quality, requiring some intervention or adjustment.	4
5		Moderate impact on production, causing some delays or noticeable quality variation.	5
6		Noticeable reduction in product quality or process efficiency, requiring significant rework.	6
7	High	Major disruptions in production or quality issues affecting customer satisfaction.	7
8	Major	Production halted for an extended period or significant batch quality loss.	8
9	Critical	Serious operational disruption, regulatory issues, or safety concerns.	9
10	Catastrophic	Complete system failure, severe safety risk, or major regulatory compliance breach.	10

Source: Authors, (2026).

Table 2: FMECA scale to rate occurrence (O).

Occurrenc e Level	Category	Description	Rank
1	Rare	Failure is extremely unlikely; no recorded incidents during plant operation.	1
2		Failure is highly unlikely; happens less than once in multiple years.	2
3	Low	Failure is infrequent; happens once every few years.	3
4		Failure is somewhat infrequent; happens about once per year.	4
5	Moderate	Failure is occasional; happens a few times per year.	5
6		Failure is somewhat frequent; happens about once a month.	6
7	High	Failure is frequent; happens multiple times per month.	7
8		Failure is very frequent; happens weekly.	8
9	Extreme	Failure occurs almost daily during operations.	9
10		Failure is nearly inevitable; happens multiple times per day without intervention.	10

Source: Authors, (2026).

Table 3: FMECA scale to rate detection (D).

Detection Level	Category	Description	Rank
1	Almost Certain	Failure is nearly always detected with current controls before it causes an issue.	1
2		Failure is very likely to be detected before causing significant effects.	2
3	High	Failure is likely to be detected under most conditions.	3
4		Failure is detected under specific conditions, but not consistently.	4
5	Moderate	Failure is occasionally detected; there is a moderate likelihood of detection before significant effects.	5

Source: Authors, (2026).

To minimize the uncertainty in assigning scales, multiple experts are involved in the process. Risk factor ranking data for all systems, components, and failure modes is gathered from multiple experts based on their years of experience, educational qualifications, quality audits undertaken, and the number of employees they have supervised as presented in Table 4.

The relative and overall weight values for the selected parameters have been calculated based on the weight values assigned to each expert. The overall weight values calculated are presented in Table 5.

3. Standardize the collected data and create the initial decision matrix 'A'. The normalized ranks for each failure mode are determined by combining the experts ranks with their corresponding overall weight values through multiplication. The various failure modes and causes of 14 systems are studied. A decision matrix is constructed for the failure modes of all components using these normalized ranks.

4. Normalize the values collected from industry experts by assigning weights to experts. The normalization process in the integrated Expert Judgment–Preferential Selection Index (ExJ–PSI) model is based on the linear normalization method outlined by [19]. After constructing the initial decision matrix using expert-weighted scores for each failure mode against criteria such as Severity, Occurrence, and Detection, the data are normalized to a common scale for comparability. This ensures that all risk factors are scaled consistently between 0 and 1, allowing for objective comparison across criteria and preventing any single factor from disproportionately influencing the results. Assign relative weights to each expert to reflect their influence on the final decision. The weight for each expert is based on the parameters listed above. These weights are normalized to derive overall expert weights. These weights are then used to adjust the expert scores in the subsequent steps.

5. Construct the initial decision matrix by considering failure mode and risk criteria. Combine expert scores (weighted by their expertise) for each failure mode and criterion to construct the initial decision matrix A. Specifically, expert ratings for each failure mode and risk criterion (Severity, Occurrence, and Detection) were collected using a standardized scale (typically 1–10). These individual expert scores were then weighted according to each expert's qualifications, experience, and relevance to the study domain (as detailed in Table 4).

Table 4: Scale assign to expert weightage.

Sr. No	Scale assign to Experts	Experience in Years	Educational Background	Quality Audits Faced	No. of Employees Supervised	Scale
1	Novice (N)	<5	Matric passed	Faced less than 2 external audits	0-10	1
2	Competent (C)	5-10	ITI or Equivalent course in Food Safety/ Mechanical Engineering	Faced 2-4 external audits	10-20	2
3	Proficient (P)	10-15	Diploma in Food Safety/ Mechanical Engineering	Successfully faced 5–6 external audits with satisfactory feedback.	20-30	3
4	Expert (E)	15-20	Bachelor Degree in Food Safety/ Mechanical Engineering	Successfully faced 7–8 external audits with satisfactory feedback.	30-40	4
5	Master (M)	>20	Masters/ PhD in Food Safety/ Mechanical Engineering	Successfully faced more than 8- 10 external audits with excellent feedback.	40-50	5

Source: Authors, (2026).

Table 5: Scale overall expert weightage.

Expert No.	Experience in years		Employee under supervision		No. of Audits faced		Educational Background			Overall Weightag e
	Weighta ge	Relative weight	Weighta ge	Relative weight	Weigh-tage	Relative weight	Qualificatio n	Weight age	Relative Weight	
Exp-1	4	0.1053	5	0.2273	5	0.1515	M.Tech	5	0.1613	0.1613
Exp-2	5	0.1316	3	0.1364	5	0.1515	B.E Instrumen-tation	4	0.1290	0.1371
Exp-3	3	0.0789	5	0.2273	5	0.1515	B.Tech Dairy	4	0.1290	0.1467
Exp-4	3	0.0789	1	0.0455	5	0.1515	B.E Electronics	4	0.1290	0.1012
Exp-5	5	0.1316	3	0.1364	5	0.1515	Diploma	3	0.0968	0.1291
Exp-6	1	0.0263	1	0.0455	4	0.1212	Diploma	3	0.0968	0.0724
Exp-7	2	0.0526	1	0.0455	1	0.0303	ITI	2	0.0645	0.0482
Exp-8	5	0.1316	1	0.0455	1	0.0303	ITI	2	0.0645	0.0680
Exp-9	5	0.1316	1	0.0455	1	0.0303	ITI	2	0.0645	0.0680
Exp-10	5	0.1316	1	0.0455	1	0.0303	ITI	2	0.0645	0.0680

Source: Authors, (2026).

The weighted average of these scores was calculated for each failure mode and criterion, resulting in the final numerical values aija_{ij}aij used in the decision matrix (presented in Table 5). This process ensures a more objective aggregation of expert judgment and is now clearly detailed in the revised manuscript for transparency and reproducibility.

$$A = a_{ij} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad (1)$$

Where aij is the index value. i=1,2,……,m along the row represents the failure modes and j=1,2,..... n represents the risk criteria's along the column.

6. Normalize decision matrix. The values are transformed in the matrix into a standardized scale so that all criteria are comparable. This step is necessary because different criteria often have different units or scales, and normalization ensures that no single criterion disproportionately influences the results.

For Maximization Problems

$$N_{ij} = \frac{a_{ij}}{a_{max}} \quad (2)$$

For Minimization Problems

$$N_{ij} = \frac{a_{ijmin}}{a_{ij}} \quad (3)$$

$$A1 = [a_{ij}] = \begin{bmatrix} N_{11} & N_{12} & \cdots & N_{1n} \\ N_{21} & N_{22} & \cdots & N_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ N_{m1} & N_{m2} & & N_{mn} \end{bmatrix} \quad (4)$$

7. Calculate the mean value of the normalized data

$$N = \frac{1}{m}\sum_{i=1}^{m} N_{ij} \quad (5)$$

8. Calculate the values of the variation of preferences between the values of every attribute

$$\theta = \frac{1}{m}\sum_{i=1}^{m} (N_{ij} - N)^2 \quad (6)$$

9. Calculate the deviations in the preference value for all criteria

$$\Omega j = (1 - \theta j) \quad (7)$$

10. Calculate the overall criteria weights for each criterion

$$w_j = \frac{\Omega_j}{\sum_{j=1}^{n} \Omega_j} \quad (8)$$

11. Prepare a matrix by multiplying Nij and ωj
12. Calculate the Criticality Index for all failure modes and all components.
13. Criticality rankings are given according to the value of CIPSI. When using Expert Judgment and the PSI model for ranking criticality, minimizing attributes during normalization involves ascending order, prioritizing failure modes with the lowest criticality index values to reduce minor risks. Conversely, maximizing attributes during normalization uses descending order, focusing on the highest values to address the most critical risks first. This ensures efficient risk mitigation and resource allocation.

Table 6 presents the various systems in Mozzarella cheese plant. The three most critical systems are assessed using the Failure Mode and Effect Analysis (FMEA) method, analyzed through the integrated Ex-PSI model. The process for determining the sequential ranking of the most critical components within these systems is conducted using the integrated Ex-PSI model. The detailed steps are explained in the later part of the section.

Table 6: Critical systems evaluated by integrated ExJ-PSI model.

Sr. No.	System	S	O	D	Criticality$_{PSI}$	Rank
1	Pre-Cooking & Cooking Stretching Machine	6.0212	5.0212	2.6231	0.7136	2
2	Stretcher Cooker	5.6470	4.5505	2.7279	0.7507	5
3	Stretching Machine	5.2727	4.6470	3.4764	0.7178	3
4	Pakonna Machine	4.7279	5.1679	2.4985	0.7732	6
5	Moulder	6.3955	5.2488	3.4764	0.6383	1
6	Brine Tank and Circulation Pump	4.0212	4.0212	3.4764	0.8432	9
7	Finishing O Vat	4.5573	4.3925	2.6231	0.8243	8
8	Tumbler	3.8745	4.3536	1.8746	0.9580	14
9	Dicer Machine	4.3536	5.5003	1.8746	0.8500	12
10	Hassia Machine	5.1021	5.6470	2.2488	0.7526	4
11	Thermoforming Packaging Machine	4.0182	4.2878	2.9555	0.8473	13
12	Nitrogen Plant	4.1830	4.7668	3.2069	0.7856	7
13	Dicer Dispenser	3.8356	4.7668	3.1021	0.8188	10
14	Double O Vat	4.2878	3.9135	3.6859	0.8250	11

Source: Authors, (2026).

The analysis showed that there were three significant risk-critical systems on the production line, namely: (1) Moulder, (2) Pre-Cooking and Cooking Stretching Machine, and (3) Stretching Machine. Once the expert-based decision matrix was constructed, normalization was done to standardize expert assessment on all criteria. The difference in preference with each criterion was then measured by calculating the variance, which is the level of disagreement on the ratings by the experts. These amounts of variance were then transformed into deviation scores ( ), representing the variation in the expert judgement. These deviations are used to calculate the complementary agreement measure (Ω) which is a measure of strength of agreement amongst experts.

The measures of agreement were standardized to produce criterion weights (wj) with attributes that had greater expert consensus having greater influence on the decision result. These weights were subsequently multiplied with the normalized failure mode scores (Nij) to calculate the final Criticality Index with the PSI-based formulation (CIPSI). This will mean that the more attributes of failure are agreed upon, the more ranking power they will have and thus enhance the reliability of prioritization. With this combined weighting and aggregation, the elements of all the three critical systems were ranked in an organized manner in terms of criticality of their failure. The result gives a systematic foundation to maintenance prioritization and reliability-oriented decision planning in mozzarella cheese processing facilities.

Table 7: Decision Matrix for PSI.

System	Component	Failure Order	Failure Mode	Failure Cause	Failure Effect	SV	O	D
						Severity	Occurrence	Detection
Moulder	Auger Gearbox	F1	Misalignment	Improper installation, deflection of shaft under load	Wear and tear, uneven wear and force distribution leading to premature failure	1.8746	3.0602	2.3536
		F2	Shaft Failure	Misalignment, Excess load, noise	Wear and tear of gearbox component	5.9974	4.9793	4.1650
		F3	Overheating	Improper lubrication	Gearbox and bearing damaged	3.8171	3.3380	4.4428
	Auger Motor	F4	Seal Failure	Oil seal leakage	Leakage in motor	3.6470	3.9165	4.2069
		F5	Deteroration of the bearing	Misalignment, Improper lubrication, Abnormal load	Excess friction and overheating, motor noise	5.9974	5.3619	5.0685
		F6	Electric Failure	Cable Short, motor trip	Motor fail	4.4613	3.6859	5.5154
	Chilled Water Pump	F7	Deteroration of the bearing	Misalignment, Improper lubrication, Abnormal load	Pump Failure	5.5184	3.5422	3.8326
		F8	Electric Failure of pump	The voltage supply to the pump is not correct	Failure of circuit breaker and winding burn	2.7279	2.1021	6.0364

		F9	nstrumentation Failure	PLC card problem, Water inside level switch	Pump Failure	4.9135	3.5393	6.7841
	Ejector Plate	F10	Mechanical failures	Plate bend, stud broken, Support rod bend	Moulder failure	4.9793	4.7098	3.8326
		F11	nstrumentation Failure	Sensor cable, Sensor springs	Plate failure	5.1830	2.9555	5.4526
Pre-Cooking & Cooking Stretchin g Machine	Feeding System	F12	Mechanical Failure	Wear and tear, material fatigue, lack of maintenance, or overloading.	Complete stoppage of curd feeding, reduced throughput, and production delays	3.9374	4.9793	5.4583
		F13	Motor or Drive System Failure	Electrical issues, mechanical overload, or motor burnout.	Inability to move curd into the machine, production halts, and potential damage to other components.	5.4583	3.5003	3.7698
	e-Cooking Section	F14	Temperature Sensor Malfunction	Sensor damage, calibration issues, or electrical faults.	Incorrect temperature readings, leading to inadequate or excessive heating	2.7279	5.0212	5.0212
	mperature Control System	F15	Sensor Malfunction	Sensor degradation, electrical issues	Incorrect temperature control leading to poor cheese quality	3.8919	3.0567	3.4729
	Water Circulation System	F16	Leakage	Worn-out seals, pipe corrosion	Loss of water pressure, potential water damage to equipment	4.5003	2.2488	3.8746
		F17	Valve Malfunction	Mechanical wear, blockage	Inability to control water flow accurately	3.3536	3.8356	4.1650
	Motor and Drive System	F18	Electrical Faults	Aging electrical components, power surges	Motor failure, potential fire hazards	5.2488	4.8746	6.3536
Stretchin g machine	Motor and Drive System	F19	Gearbox Failure	Insufficient lubrication, mechanical overload	Reduced efficiency, potential stoppage	6.9944	1.7279	6.1411
	Water Circulation System	F20	Pump Failure	Mechanical wear, electrical issues	Insufficient water flow, overheating of the product	5.8167	4.7698	5.8710
		F21	Leaks	Damaged pipes, loose connections, or faulty seals	Loss of water, reduced efficiency of the cooling system	2.9165	3.9165	5.7279

Source: Authors, (2026).

Normalized ranks are evaluated for all failure modes by considering experts overall weights. Decision matrix is created for all the components of identified most critical systems. Table 7 shows the decision matrix of three most critical systems i.e., for Moulder, Pre- Cooking & Cooking Stretching Machine and Stretching machine in mozzarella cheese processing plant. The normalized value of decision matrix is performed using above equation and the results are presented in Table 8. In multi-attribute decision-making methods, it is essential for attribute values to be dimensionless. This is accomplished by scaling each attribute to a range of 0 to 1, a process referred to as normalization. Moreover, all three parameters in this analysis are considered minimization criteria.

Table 8: Normalized decision matrix for PSI.

System	Component	Failure Order	SV	O	D
			Severity	Occurrence	Detection
Moulder	Auger Gearbox	F1	1	0.5646	1
		F2	0.3126	0.3470	0.5651
		F3	0.4911	0.5176	0.5298
	Auger Motor	F4	0.5140	0.4412	0.5595
		F5	0.3126	0.3223	0.4644
		F6	0.4202	0.4688	0.4267
	Chilled Water Pump	F7	0.3397	0.4878	0.6141
		F8	0.6872	0.8220	0.3899
		F9	0.3815	0.4882	0.3469
	Ejector Plate	F10	0.3765	0.3669	0.6141
		F11	0.3617	0.5846	0.4316
Pre-Cooking & Cooking Stretching Machine	Feeding System	F12	0.4761	0.3470	0.4312
		F13	0.3434	0.4936	0.6243
	Pre-Cooking Section	F14	0.6872	0.3441	0.4687
	Temperature Control System	F15	0.4817	0.5653	0.6777
	Water Circulation System	F16	0.4165	0.7683	0.6074
		F17	0.5590	0.4505	0.5651
	Motor and Drive System	F18	0.3571	0.3545	0.3704
Stretching machine	Motor and Drive System	F19	0.2680	1	0.3833
	Water Circulation System	F20	0.3223	0.3623	0.4009
		F21	0.6427	0.4412	0.4109

Source: Authors, (2026).

The average values of the normalized data for each risk factor are computed.

NSV = 0.4643, NO = 0.5017, ND = 0.51819

The values of the variations of preference are calculated

θ SV = 0.6064, θ O = 0.6127, θ D = 0.4413

The values of the deviation of variations of preference are calculated

Ω SV = 0.3935, Ω O = 0.3872, Ω D = 0.5586

The overall criteria weights for all risk factors are calculated

ωSV = 0.2938, ωO = 0.2890. ωD = 0.4171

After calculating overall weights, the result is obtained by preparing a matrix by multiplying Nij and ωj. Table 9 shows the value obtained by multiplying Nij and ωj.

Table 9: Multiplication matrix of Nij and ωj.

System	Component	Failure Order	SV	O	D
			Nij x ωj	Nij x ωj	Nij x ωj
Moulder	Auger Gearbox	F1	0.2931	0.1632	0.5587
		F2	0.0918	0.1003	0.3157
		F3	0.1443	0.1496	0.2960
	Auger Motor	F4	0.1510	0.1275	0.3126
		F5	0.0918	0.0932	0.2594
		F6	0.1234	0.1355	0.2384
	Chilled Water Pump	F7	0.0998	0.1410	0.3431
		F8	0.2019	0.2376	0.2178
		F9	0.1121	0.1411	0.1938
	Ejector Plate	F10	0.1106	0.1061	0.3431
		F11	0.1063	0.1690	0.2412
Pre- Cooking & Cooking Stretching Machine	Feeding System	F12	0.1399	0.1003	0.2409
		F13	0.1009	0.1427	0.3488
	Pre-Cooking Section	F14	0.2019	0.0995	0.2619
	Temperature Control System	F15	0.1415	0.1634	0.3786
	Water Circulation System	F16	0.1224	0.2221	0.3394
		F17	0.1642	0.1302	0.3157
	Motor and Drive System	18	0.1049	0.1025	0.2070
Stretching machine	Motor and Drive System	F19	0.0787	0.2891	0.2141
	Water Circulation System	F20	0.0947	0.1047	0.2240
		F21	0.1888	0.1275	0.2296

Source: Authors, (2026).

The values of the criticality index of each failure mode are obtained. Finally, the criticality index of a component is calculated by adding the values of the criticality index of all its failure modes. The same analysis was performed on all the 21 failure modes of the three critical systems of the mozzarella system in Table 10. The ranking of the components was then done according to their criticality index values and the findings of all the 11 chosen components in the three critical systems were tabulated in Table 12. Although several root causes were determined to cause the few failure modes, only the most important ones were taken into account in this study. Also, the values of the criticality index and their respective ranks were determined on some selected mozzarella systems.

The evaluation of the component criticality indices (CI) introduced in Table 11 allows identifying the most important systems that should be given special attention. The Pre-Cooking and Cooking Stretching Machine (CI: 3.2936) is the most important system, then there are the Moulder (2.3615) and the Stretching Machine (1.1231). In the Pre-Cooking and Cooking Stretching Machine, the most important component is the Temperature Control System (0.7686), and therefore predictive maintenance is necessary. The Moulder needs to attend to its Chilled Water Pump (0.6431) and Auger Motor (0.5732) whereas the Stretching Machine needs to look at the Motor and Drive System (0.6422). An organized maintenance plan according to these rankings can lead to better reliability, fewer failures and better operation efficiency.

Table 10: Critical ranks for failure modes.

System	Component	Failure Order	Failure Mode	Rank	Overall CI of Component	Overall CI of Sub- System (Failure modes)	Criticality Rank
Moulder	Auger Gearbox	F1	Misalignment	21	0.8742	5.5694	1
		F2	Shaft Failure	5	0.4278		
		F3	Overheating	15	0.5149		
	Auger Motor	F4	Seal Failure	14	0.5119		
		F5	Deterioration of the bearing	3	0.3787		
		F6	Electric Failure	7	0.4370		
	Chilled Water Pump	F7	Deterioration of the bearing	12	0.4969		
		F8	Electric Failure of pump	20	0.6022		
		F9	Instrumentation Failure	4	0.3979		
	Ejector Plate	F10	Mechanical failures	11	0.4728		
		F11	Instrumentation Failure	9	0.4553		
Pre-Cooking & Cooking	Feeding System	F12	Mechanical Failure	6	0.4201	3.4984	2
		F13	Motor or Drive System Failure	13	0.5040		
	Pre-Cooking Section	F14	Temperature Sensor Malfunction	7	0.4469		
Stretching Machine	Temperature Control System	F15	Sensor Malfunction	18	0.5876		
	Water Circulation System	F16	Leakage	19	0.5978		
		F17	Valve Malfunction	16	0.5302		
	Motor and Drive System	F18	Electrical Faults	2	0.3619		
Stretching machine	Motor and Drive System	F19	Gearbox Failure	17	0.5276	1.3820	3
	Water Circulation System	F20	Pump Failure	1	0.3666		
		F21	Leaks	10	0.4878		

Source: Authors, (2026).

The overall criticality indicator for each subsystem is calculated by aggregating the Criticality Index values of all individual failure modes within that subsystem. Specifically, a weighted summation approach is applied, where each failure mode's Criticality Index is multiplied by its respective weight—reflecting its relative importance or impact—and then summed to obtain the subsystem's overall Criticality Index. This method ensures that more significant failure modes contribute proportionally more to the subsystem's total criticality, providing a comprehensive view of system risk.

Table 11: Critical ranks for prominent critical systems.

System	Component	Overall CI of Component	Overall CI of Sub-System	Criticality Rank
Moulder	Auger Gearbox	0.5604	2.3615	2
	Auger Motor	0.5732		
	Chilled Water Pump	0.6431		
	Ejector Plate	0.5848		
Pre- Cooking & Cooking Stretching Machine	Feeding System	0.6522	3.2936	1
	Pre-Cooking Section	0.6851		
	Temperature Control System	0.7686		
	Water Circulation System	0.7091		
	Motor and Drive System	0.4787		
Stretching machine	Motor and Drive System	0.6422	1.1231	3
	Water Circulation System	0.4810		

Source: Authors, (2026).

Table 12: Comparative analysis of results for components of three critical subsystems.

System	Component	Rank by proposed model	Conventional RPN method	Normalized median
Moulder	Auger Gearbox	3	3	3
	Auger Motor	4	4	4
	Chilled Water Pump	7	8	6
	Ejector Plate	5	6	5
Pre-Cooking & Cooking Stretching Machine	Feeding System	8	9	7
	Pre-Cooking Section	9	5	9
	Temperature Control System	11	11	11
	Water Circulation System	10	10	10
	Motor and Drive System	1	1	1
Stretching machine	Motor and Drive System	6	7	8
	Water Circulation System	2	2	2

Source: Authors, (2026).

Comparison of results was done by using the Proposed Model, the Conventional RPN method, and the Normalized Median approach. The Normalized Median Rank provides a standardized measure that helps in comparing the criticality of failure modes across different systems or analyses. It aids in making informed decisions about where to focus resources for risk reduction and reliability improvement. The final results for the systems of the mozzarella cheese plant by the proposed integrated model (ExJ-PSI), conventional FMECA method and normalized median method is presented in Table 13.

Table 13: Comparative analysis of results for all subsystems.

Sr. No.	System	Rank by proposed model	Conventional RPN method	Normalized Median
	System	ExJ-PSI	Conventional RPN Rank	Normalized Median Rank
1	Moulder	1	1	1
2	Pre-Cooking & Cooking Stretching Machine	2	3	3
3	Stretching Machine	3	2	2
4	Stretcher Cooker	4	6	4
5	Hassia Machine	5	7	5
6	Pakonna Machine	6	8	6
7	Nitrogen Plant	7	4	7
8	Dicer Dispenser	8	12	9
9	Finishing O Vat	9	9	10
10	Double O Vat	10	5	8
11	Brine Tank and Circulation Pump	11	11	11
12	Thermoforming Packaging Machine	12	10	12
13	Dicer Machine	13	13	13
14	Tumbler	14	14	14

Source: Authors, (2026).

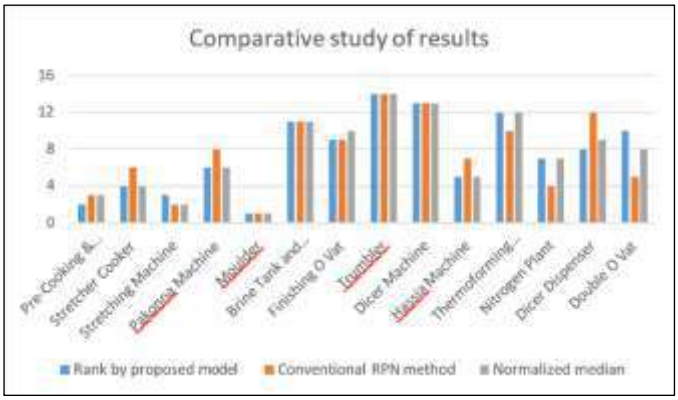


Figure 2: Comparative result of ranking of systems in Mozzarella cheese system.
Source: Authors, (2026).

The comparison between the rankings obtained with the proposed model, the traditional RPN method, and the normalized median values indicates that all the three methods determine the same key systems, summarized in Table 13. The ExJ-PSI method is more accurate with expert assigned weights which restricts the bias that is usually observed in conventional RPN scoring. Its normalization process also facilitates the comparison of various criteria on a similar scale that creates more definite and dependable risk priorities.

The approach also enhances efficiency in that it integrates the expertise with quantitative data in an organized manner, enabling the decision-makers to allocate resources faster and with more confidence. The ExJ-PSI model in practice is effective in complex settings like the mozzarella cheese processing where a combination of risk factors plays out. It is also very handy to both maintenance and quality teams due to its flexibility. In all three ranking techniques, the findings are similar.

The Moulder system is consistently ranked at Rank 1 which proves its significance in the entire process. The Stretching Machine remains on Rank 2 and the Pre-Cooking and Stretching Machine still remains at Rank 3. These consistent positions imply that such systems should be given priority to ensure a smooth and dependable running of plants. Minor variations in the rankings are observed when the normalized median values are compared with the suggested ExJ-PSI model and the classical RPN method. As an example, the Stretcher Cooker is ranked 4th in the proposed model and the normalized median approach, but 6th in the RPN approach, which indicates that its real risk could be more than the traditional RPN approach suggests. The same can be observed in the case of the Pakonna Machine, which is ranked 6th in the proposed and normalized ranking, but on the RPN one, the machine is ranked 8th, which indicates that there is a potential to misclassify the risk level.

Other statistically significant deviations are the Dicer Dispenser, which the proposed model ranks 8 th, the RPN method ranks 12 th and the normalized median method ranks 9 th. This diffusion means that the traditional RPN approach might underestimate its risk. The Nitrogen Plant is also inconsistent, keeping 7th position in the proposed and normalized approaches but increasing to 4th in the RPN method, indicating that statistical normalization offers a more consistent and realistic prioritization. The rankings of the Brine Tank and Circulation Pump (Rank 11), the Tumbler (Rank 14), and the Dicer Machine (Rank 13) have the same rankings in all three methods of evaluation, and there is a definite consensus that these systems are relatively less risky. The fact that the proposed model is close to the traditional RPN approach also contributes to the credibility of the new method. Simultaneously, the slight changes observed in the normalized median rankings suggest that statistical normalization can be added to identify subtleties that would otherwise be missed to provide the maintenance and risk-management choices with greater precision.

In comparison with FMEA and FMECA, the ExJ-PSI approach offers better grounds to prioritize failure modes within the setting like mozzarella cheese production. The conventional FMEA tends to generate duplicative RPN scores and does not distinguish between them, but ExJ-PSI uses expert-established weights and normalized factors to provide more consistent and less biased scores. FMECA is an improvement of FMEA with an element of criticality, yet it does not provide the flexibility and decision-making that are integrated in ExJ-PSI. With a combination of professional judgment and numerical scoring, ExJ-PSI method provides a more accurate prioritization, is more appropriate to food processing systems, and enhances the alignment of maintenance, reliability, and safety objectives. These results demonstrate the significance of periodically revising and enhancing the risk-assessment instruments to enhance the system performance and reliability of operation.

Figure 3 shows the summary of the work conducted on the mozzarella cheese system. The three most important subsystems i.e. Moulder, Pre-cooking and Stretching machine and Stretching machine of mozzarella cheese system is further analysed to rank the components on the basis of criticality. The results are categorized according to the level of criticality i.e. High criticality level, Moderate criticality level and low criticality level. The components with high criticality level that are ranked are Auger gearbox, Auger motor, Ejector plate, Motor and drive system, water circulation system and the Predictive based maintenance (PdM) and Condition based maintenance (CBM) plan is proposed to these components. Moreover, moderate criticality level components ranking evaluated are Chilled water pump, Feeding system, Pre-cooking system, Motor and drive system and the preventive maintenance plan is minimal is proposed to these components. These maintenance plan would be beneficial in terms of productivity with least break down and greatly enhanced the quality of cheese and the safety of the mozzarella cheese system.

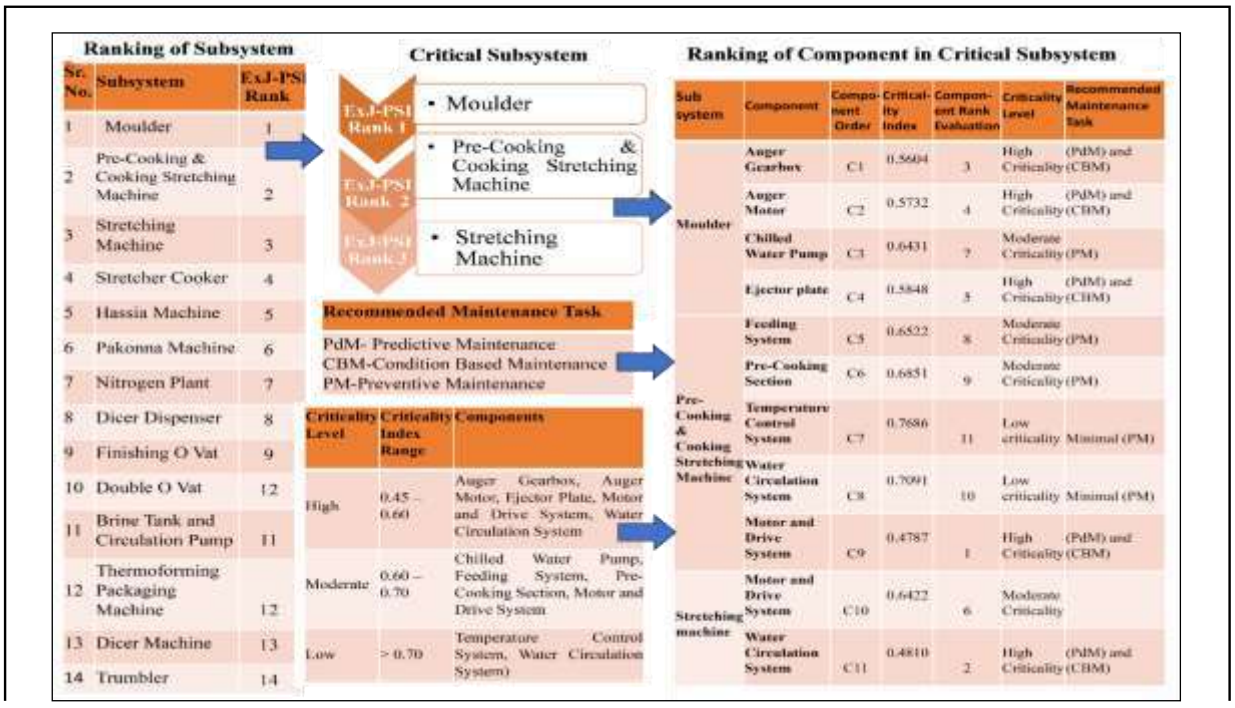


Figure 3: Summary of ranking of critical subsystem and components of critical subsystem with recommendation of maintenance task.
Source: Authors, (2026).

IV. CONCLUSIONS

According to the analysis of the critical systems of an industrial mozzarella cheese facility, the integrated Expert Judgment-Preference Selection Index (ExJ-PSI) model was an effective and systematic way of detecting and prioritizing failure criticality. The methodology has succeeded in overcoming the limitations of the traditional FMECA-based ranking, especially the subjective scoring dispersion, disparity in weighting the criteria, and the ambiguity in evaluating the importance of failure. The most important conclusions made in this piece of work are as follows:

- The ExJPSI model defines plausible and bias-managed prioritization of failure modes. The framework reduced the variability in personal judgement by normalizing the opinions of several groups of experts, and converting them into objective weights, to achieve consistency in the evaluation. This method was very successful in solving close-ranking failure modes which could not be accurately differentiated by traditional RPN-based methods. The high congruence of the rankings obtained by ExJ-PSI and traditional procedures confirms its utility in the industrial environment and makes maintenance choices consistent with the operational risks that are realistic.
- Three systems were determined to be production-critical because of direct impact on the product formation and downstream reliability. The Pre-Cooking & Cooking Stretching Machine was the most important subsystem followed by the Moulder and the Stretching Machine. The research indicated that performance variability in these units has a major impact on the ultimate cheese structure, stretching behaviour, distribution of moisture and uniformity of the product. These results point to the necessity to distribute maintenance activities in a top-down fashion, starting with the subsystem that is most at risk.
- The component-level classification is a systematic maintenance roadmap to improve reliability. The critical components such as auger gearbox, auger motor, ejector assembly, and drive system need predictive or condition based maintenance to avoid failure. Items that are medium critical like water pumps, feeders, and pre-cooking units can be well handled by scheduled preventive servicing. Components with low criticality require simple routine checks. This classification also allows manufacturers to streamline the stocking of spare parts, maintenance schedules and staffing.
- The findings indicate that there are quantifiable industrial gains of implementing the ExJ-PSI-based ranking process. The strategy aids in minimizing unplanned stoppages, increased consistency in processes, greater use of assets, and greater forecasting of maintenance workload. Therefore, the ExJ-PSI model provides a scalable decision-support framework that can be generalized to other dairy-based processing systems and other multi-stage industrial equipment that needs the focus on reliability.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Vaishali Talkit, Anand Bewoor and Amitkumar Shelar.
Methodology: Vaishali Talkit and Anand Bewoor.
Investigation: Vaishali Talkit and Amitkumar Shelar.
Discussion of Results: Vaishali Talkit, Anand Bewoor and Amitkumar Shelar.
Writing – Original Draft: Vaishali Talkit.
Writing – Review and Editing: Vaishali Talkit and Anand Bewoor.
Resources: Anand Bewoor.
Supervision: Anand Bewoor and Amitkumar Shelar.
Approval of the Final Text: Vaishali Talkit, Anand Bewoor and Amitkumar Shelar.

VI. REFERENCES

[1] Pereira, H. C. (1993). Food production and population growth. Land use policy, 10(3), 187-190. https://doi.org/10.1016/0264-8377(93)90013-Z

[2] Igbokwe, N. C., & Godwin, H. C. (2021). Maintenance performance evaluation and downtime analysis of manufacturing equipment in a food manufacturing company. Journal of Engineering Research and Reports, 100-107. https://dx.doi.org/10.9734/jerr/2021/v20i1117413

[3] Gedikli, T., & Cayir Ervural, B. (2019, September). Selection of optimum maintenance strategy using multi-criteria decision making approaches. In Global Joint Conference on Industrial Engineering and Its Application Areas (pp. 156-170). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030- 42416-9_15

[4] Soltanali, H., Khojastehpour, M., & Torres Farinha, J. (2023). An improved risk and reliability framework-based maintenance planning for food processing systems. Quality Technology & Quantitative Management, 20(2), 256-278. https://doi.org/10.1080/16843703.2022.2093565

[5] SZCZYRBA, A., & DZIUBA, S. (2023). Good Manufacturing Practices for Quality and Safety Management in the Food Industry. Materials Research Proceedings, 34. https://doi.org/10.21741/9781644902691-34

[6] Tsarouhas, P. H. (2011). A comparative study of performance evaluation based on field failure data for food production lines. Journal of Quality in Maintenance Engineering, 17(1), 26-39. https://doi.org/10.1108/13552511111116231

[7] Rahardjo, B., & Jiang, B. C. (2016). An analysis for providing safety in the cooking oil production process through FMECA approach. International Journal of Basic and Applied Sciences, 5(2), 151-156.

[8] Siekelova, A., Podhorska, I., & Imppola, J. J. (2021). Analytic hierarchy process in multiple–criteria decision–making: a model example. In SHS web of conferences (Vol. 90, p. 01019). EDP Sciences. https://doi.org/10.1051/shsconf/20219001019

[9] Ansah, R. H., Sorooshian, S., & Bin Mustafa, S. (2015). Analytic hierarchy process decision making algorithm. Global journal of pure and applied mathematics, 11(4).

[10] Madanchian, M., & Taherdoost, H. (2023). A comprehensive guide to the TOPSIS method for multi-criteria decision making. Madanchian M, Taherdoost H. A comprehensive guide to the TOPSIS method for multi-criteria decision making. Sustainable Social Development, 1(1), 2220. http://doi.org/10.54517/ssd.v1i1.2220

[11] Roszkowska, E. (2011). Multi-criteria decision making models by applying the TOPSIS method to crisp and interval data. Multiple Criteria Decision Making/University of Economics in Katowice, 6(1), 200-230.

[12] Harahap, M. A. K., Tarigan, I. M., Kraugusteeliana, K., Simanullang, S., Avila, E. C., & Rahim, R. (2023). A multi-criteria analysis of university housing options using weighted sum model. JINAV: Journal of Information and Visualization, 4(1), 93-98. https://doi.org/10.35877/454RI.jinav1697

[13] Hajiagha, S. H. R., Hashemi, S. S., Mohammadi, Y., & Zavadskas, E. K. (2016). Fuzzy belief structure based VIKOR method: an application for ranking delay causes of Tehran metro system by FMEA criteria. Transport, 31(1), 108-118. https://doi.org/10.3846/16484142.2016.1133454

[14] Petković, D., Madić, M., Radovanović, M., & Gečevska, V. (2017). Application of the performance selection index method for solving machining MCDM problems. Facta Universitatis, Series: Mechanical Engineering, 15(1), 97-106. http://doi.org/10.22190/FUME151120001P

[15] Talkit, V. G., Honrao, P., Bhosale, S., Bewoor, A., & Patil, R. B. (2023). Application of the reliability analysis for modifications in maintenance scheduling. Materials Today: Proceedings, 77, 579-585. https://doi.org/10.1016/j.matpr.2022.10.075

[16] Talkit, V., Kininge, R., Kokate, P., Narkhede, D., Madame, S., Bewoor, A., & Patil, R. B. (2023). Reliability analysis of repairable and replaceable system: Dairy product industry. Materials Today: Proceedings, 77, 573-578. https://doi.org/10.1016/j.matpr.2022.10.074

[17] Patil, S. S., & Bewoor, A. K. (2021). Reliability analysis of a steam boiler system by expert judgment method and best-fit failure model method: a new approach. International Journal of Quality & Reliability Management, 38(1), 389-409. https://doi.org/10.1108/IJQRM-01-2020-0023

[18] Patil, S. S., Bewoor, A. K., Patil, R. B., Kumar, R., Ongar, B., Sarsenbayev, Y., ... & Elsheikh, A. (2022). A new approach for failure modes, effects, and criticality analysis using ExJ-PSI model—a case study on boiler system. Applied Sciences, 12(22), 11419. https://doi.org/10.3390/app122211419

[19] Pancholi, N., & Bhatt, M. (2018). FMECA-based maintenance planning through COPRAS-G and PSI. Journal of Quality in Maintenance Engineering, 24(2), 224-243. https://doi.org/10.1108/JQME-03-2017-0015

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 846-855. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3094

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

SCIENTIFIC RESEARCH AS A KEY ELEMENT OF INSTITUTIONAL TRANSFORMATION IN THE PERUVIAN ARMY: A SYSTEMATIC REVIEW

Michael Joaquin Herrera Contreras[1], Aaron Aguirre Venturi*[2] and Ada Gallegos[3]*

[1, 2]Instituto Científico Tecnológico del Ejército, Peru.
[3]Universidad Nacional Mayor de San Marcos, Peru.

[1]https://orcid.org/0000-0002-6999-538X, [2]https://orcid.org/0000-0001-9242-6410, [3]https://orcid.org/0000-0002-8264-711X

Email: *agallegosr@unmsm.edu.pe, aaronaguirreventuri@gmail.com, michaeljherreracontreras@gmail.com

ARTICLE INFO

Article History
Received: December 9, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
research,
science and development,
science administration,
research and development,

ABSTRACT

The objective of this study was to determine the contribution of scientific research to institutional transformation in the Peruvian Army. The eligibility criteria applied were topic, sources, methodological design, related topics, language, and publication date, which allowed for the analysis of 212 studies from different academic databases. Exclusion criteria, in a first phase, excluded duplicate studies; in a second phase, those that covered study topics and variables unrelated to the research; and in a third phase, those that did not provide the required information, after analyzing the title, abstract, and full text. The databases reviewed were Scopus, ScienceDirect, Scielo, Redalyc, ERIC, and Dialnet, during the first days of June 2025. Fourteen studies were included, whose main characteristic is their focus on strengthening education, research, and innovation in military and academic contexts. The main results show that scientific research constitutes a fundamental activity for the transformation of an institution, by strengthening ethical and professional training, promoting critical thinking, and optimizing evidence-based decision-making, fostering a culture of continuous improvement through the use of innovative methods and the inclusive participation of diverse stakeholders. Furthermore, this analysis revealed a limited research culture, which restricts the assessment, development, and effective application of scientific research in the process of institutional transformation. It is concluded that, although scientific research plays a crucial role in the institutional transformation of the Peruvian Army, its impact is conditioned by a still nascent research culture, which limits its full appreciation and application.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

As [1] assert, "technology and innovation in the defence and armed forces sector are critical components for guaranteeing a country's security and functioning." In essence, contemporary defence systems must transcend a reliance on operational force and instead integrate scientific knowledge and innovation capacity as foundational elements. This integration is imperative for the effective anticipation of threats, the adaptation to novel scenarios, and the maintenance of a credible deterrent posture. [2] asserts that the Peruvian Army is facing a number of challenges, as outlined in the Institutional Transformation Plan (PTI), including a lack of equipment, infrastructure, and other logistical capabilities.

These deficiencies have a detrimental effect on the Army's operational capacity. It is evident that the issue at hand cannot be resolved solely through investments in infrastructure or imported equipment. Instead, there is a necessity to fortify internal capabilities for diagnosis, solution design, and continuous improvement. These processes are contingent on knowledge generation and management. [3] asserts that "scientific research is a prerequisite for progress in all areas of society, and military science is no exception." Consequently, research stemming from military experience and context addresses real-world problems and offers applicable solutions that can directly contribute to institutional modernisation, doctrinal innovation, and improved performance in complex scenarios.

Furthermore, [4] state that "military organisations have a structure and a leading role in strategic areas focused on science, technology, and open innovation, conducting strategic programs and projects of economic and social interest," which allows for the integration of scientific research with decision-making at all levels of command. The present research seeks to transform scientific research into an active instrument for continuous improvement, innovation, and institutional transformation, thus closing the gap between the knowledge generated and its effective use in military operations. A particularly illustrative case is the diagnosis carried out in the Colombian Air Force (FAC), where the operational leaders themselves recognise that many of their needs can be addressed internally through scientific research processes. According to [5], "The mission area agrees that a significant proportion of the personnel requirements can be met through in-house research, but the ST process must be updated to achieve this."

This statement reveals an institutional disposition towards applied research. As [6] asserts, it is imperative to analyse institutions from the standpoint of organisational culture, as it is widely acknowledged to be a pivotal element in the realisation of an organisation's objectives. In the absence of an institutional environment that acknowledges the significance of generating and applying scientific knowledge, even the most meticulously designed reforms risk being rendered ineffective. In the Peruvian Army's current educational system, there is a hierarchical structure that begins with the Ministry of Defence, passes through the General Directorate of Education and Doctrine, and continues on to the Army's Directorate of Education and Doctrine (DIEDOCE), the Army's Education and Doctrine Command (COEDE), and finally, the Army's educational institutions, which are subordinate to this command. The Army is responsible for the provision of undergraduate and postgraduate educational programmes.

The former encompasses university professional training and technical/technological professional training, while the latter comprises doctorates, master's degrees, diplomas, second professional specialisations and continuing education. The promotion and development of military research is facilitated through the academic programmes conducted by these institutions [7]. The significance of this research lies in the educational responsibility incumbent upon the Army with regard to the challenges involved in implementing an effective and structured system that fosters scientific research. These challenges encompass the inadequate management of scientific research, the necessity to increase the impact and significance of published research, and the improvement and augmentation of resources necessary for the sustainability of scientific research processes. The rationale underpinning this research is the pursuit of enhancing scientific output within the Army, a pursuit that will yield dual benefits for its members.

On the one hand, it will enable them to cultivate critical and strategic skills, and on the other, it will have a positive societal impact in Peru by generating knowledge applicable to national defence, security, and social development. Conversely, this research is pertinent to the management of scientific research in the Peruvian Army, with a particular focus on the undergraduate and graduate schools that fall under the jurisdiction of the Army Doctrine Directorate (JDOCE), within the Army Education and Doctrine Command (COEDE), as these entities are responsible for scientific production within the institution. In this sense, the general research question guiding this study is as follows: How does scientific research contribute to institutional transformation in the Peruvian Army? The specific research questions to be addressed are as follows:

This study sets out to investigate the influence of scientific research results on modernisation processes, doctrine and decision-making in the Peruvian Army. Furthermore, it seeks to ascertain the perception of members of the Peruvian Army with regard to the impact of scientific research on organisational change. The aforementioned limitations have a restrictive effect on the institutional capacity to contribute innovative knowledge that would otherwise strengthen the functions of the Peruvian Army. The objective of this study is to determine the contribution of scientific research to institutional transformation in the Peruvian Army. The objective of this research is to establish an optimal scientific research management system that allows the research carried out by the members of the institution to be used, providing tools for the improvement of the processes and development of these.

I.1 BACKGROUND

In [8] posit in their research, titled "Security and Defense, Technological Development, Innovation, and Transformation," that contemporary threats are inducing a transformation of military forces to operate in new, complex scenarios. According to [9] conducted a study with the title "Changing Mindsets: The Transformation Process of the Colombian National Army (2011–2018)," it is stated that "education, both domestically and internationally, has facilitated the conceptualisation and implementation of an institutional cognitive transformation from the top-down, with the endorsement of the Army itself for its design and execution."In the present study, [10] explore the intricacies of "Technology and Innovation Management in the Defense Sector: Results from a Bibliometric Analysis," states that "there is a necessity for structured and systematic processes for managing technology and innovation, which allow for supporting decision-making regarding R&D&I processes.

"In his research entitled "Some Ideas on Scientific Research" [11], the author asserts that the purpose of scientific research is to verify, refute, expand upon, or propose new experimental or theoretical contributions to the body of science being addressed. In his research paper, entitled "Educational Management as a Driver of Scientific and Technological Research:", By [12] explores the potential impact of educational management on the promotion of scientific and technological research. New horizons for innovation" [Educational management as a driver of scientific and technological research: New horizons for innovation] mentions, "Education and research have a reciprocal relationship. It is evident that research produces education, while education, in turn, enhances research. Both of these factors enable individuals to engage in recreational activities, enhance their abilities, develop new skills, and achieve personal growth within the designated space-time continuum.

II. METHOD

In order to ensure the rigour, transparency, and reproducibility of this study, the PRISMA methodology was employed. This methodology facilitates the systematic structuring of the identification, selection, and analysis of relevant scientific literature. The PRISMA statement, published in 2009, is a reporting guideline designed to address the problems in the publication of systematic reviews [13].

II.1 ELIGIBILITY CRITERIA.

The following eligibility criteria were established for this research:

- Topic: Research that links scientific research as a driver of institutional transformation in the Army or armed forces of Peru or other countries.
- -Literature sources: Peer-reviewed scientific articles published in indexed academic journals.
- -Methodological design: Studies may employ a quantitative, qualitative, or mixed-methods design.
- -Related topics: Research addressing concepts related to scientific research, scientific study, and academic research, along with terms associated with the military sphere, such as army and armed forces, is considered.
- -Language: Studies must be written in English, Portuguese, and Spanish to facilitate analysis and understanding, and to maintain a Latin American context throughout the research.
- -Publication date: Research must have been published between January 1, 2021, and May 30, 2025.

II.2 SOURCES OF INFORMATION

The following platforms were consulted in order to conduct the research: Scopus and ScienceDirect, which offer high-impact, multidisciplinary academic literature; SciELO and Redalyc, which focus on research in Spanish from Latin America and the Caribbean; ERIC, which specialises in educational studies; and Dialnet, which provides access to scientific output in Spanish, especially in the social sciences and humanities. The preceding five-year period was the subject of the study. The databases were consulted on 6 June 2025.

II.3 SEARCH STRATEGIES

For the search strategy, the keywords “scientific research” and “Army” were identified, along with their corresponding synonyms. Using the Boolean operators AND, OR, NOT, (…) and “…”, the following search equation was applied:

((“scientific research”) OR (“scientific study”) OR (“academic research”)) AND ((army) OR (“armed forces”)).

The filters applied to all databases were:
- Subject areas: Social Sciences
- Access type: Open Access
- Publication dates: January 1, 2021 to May 30, 2025

II.4 SELECTION PROCESS

The selection process will be carried out in accordance with the guidelines established by the PRISMA methodology. A systematic search strategy will be conducted in six recognised academic databases, with the objective of compiling all potentially relevant research. Subsequent to the compilation of the records, the implementation of automated tools and manual review will be employed for the purpose of the elimination of potential duplicates. Subsequent to this, the studies will be evaluated through a progressive analysis. Initially, the titles will be examined, followed by the abstracts, and finally, the full texts of the studies that have passed the previous stages. The present review will be based on a set of predefined criteria, with a focus on thematic relevance, specific variables, and methodological quality. The data cleaning process will be executed in two successive phases to ensure that the studies included in the final analysis rigorously address the study's objectives.

III. RESULTS

A total of 212 studies were identified through searches in six academic databases: Scopus (n=29), ScienceDirect (n=46), SciELO (n=12), Redalyc (n=80), ERIC (n=13), and Dialnet (n=32). Following the compilation of the results, two duplicate studies were eliminated through the utilisation of automated tools and manual review, resulting in a total of unique studies that advanced to the subsequent stage. During the selection process, 183 studies were excluded. Of these, 159 were discarded for not directly addressing the study's central theme, which was the role of scientific research in institutional transformation processes within military or similar contexts. Furthermore, 24 studies were excluded on the grounds that they failed to consider the specific variables defined in this review. These included the link between academic research, organisational change, and the armed forces environment. This filtration process was implemented to ensure that the analysis remained exclusively focused on literature that was deemed relevant to the research objectives.During a subsequent review phase, 13 further studies were excluded.

Following a thorough review of the titles, five of these were discarded as they were deemed to be unrelated to the research topic. Following a thorough review of the abstracts, a further five were disqualified on the grounds that they failed to adequately address the correlation between scientific research and institutional transformation. Following a thorough examination of the full texts of three studies, they were ultimately excluded from the review due to their failure to meet the established criteria for analytical depth and methodological relevance. The process of identifying, selecting, and excluding studies was carried out with methodological rigor. This ensured that the works ultimately included in the systematic review met the criteria for relevance, scientific quality, and thematic pertinence, thus guaranteeing a solid foundation. In order to facilitate a more comprehensive understanding of the methodology, the flowchart (Figure 1) is presented, illustrating the application of the PRISMA approach.

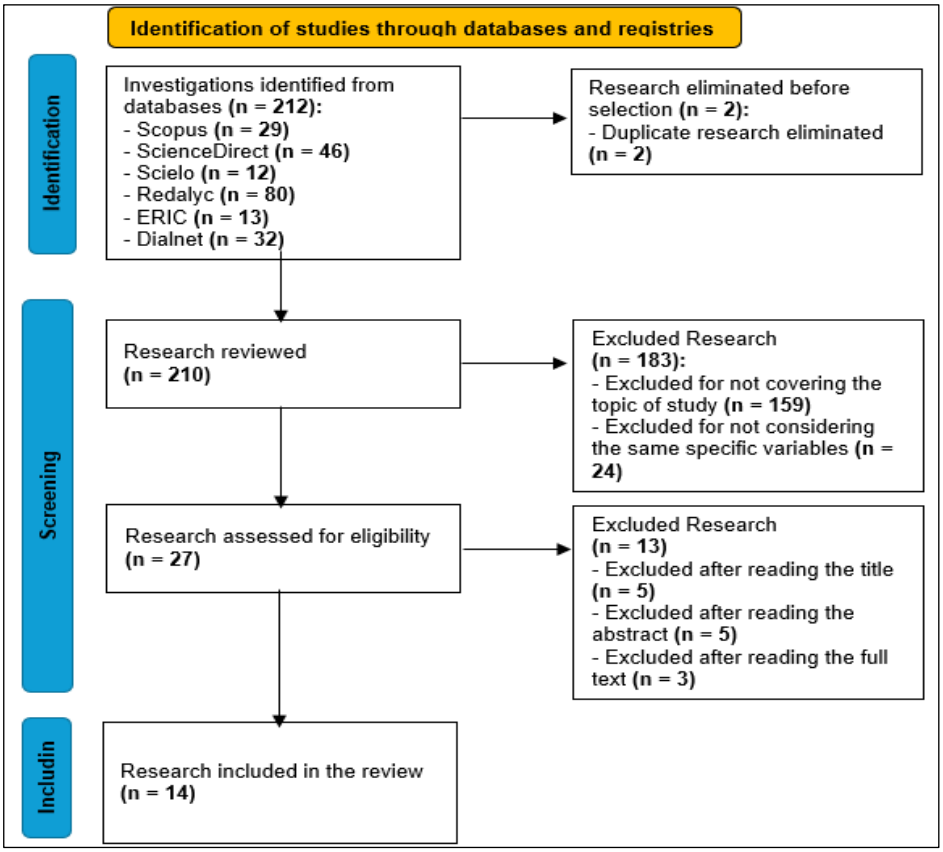


Figure 1: PRISMA flowchart for item selection.
Source: Authors, (2026).

The exclusion process was fundamental to ensuring that the final analysis was based solely on scientific literature aligned with the research objectives. The stages at which various studies were discarded, as well as the specific reasons justifying their exclusion, are detailed below. Following a thorough examination of the titles, it became apparent that the subsequent studies were to be excluded on account of the fact that they did not directly address the relationship between scientific research and institutional transformation in military contexts. Furthermore, they did not take into consideration the specific variables that had been defined for the present systematic review. In [14] published an article entitled "Improving Language Learning with Virtual Reality: A Quasi-experimental Study" focuses on language learning through immersive technologies, without any relation to organisational structures or armed forces. By [15], "Ecuadorian Air Force: On the Road to Multidomain. A Transdisciplinary Analysis of Complexity", despite its focus on a military institution, exhibits a geostrategic emphasis and does not prioritise scientific research as a transformative axis.

In their 2022 publication, [16] explore the concept of "Global and Local Knowledge for English Language Learning: A Study at a National Police School", focuses on English language learning in a police academy, without any connection to institutional transformation or the Army as the institution under study. In [17] have utilised the PLS model to explore the dynamics of HRM. The moderating roles of self-efficacy and organisational commitment in work engagement and talent turnover intention" addresses human resource dynamics in a general organisational context, without mentioning the defence sector or the use of research as a driver of institutional change. [18] provide an analysis of "Gender Education of Cadets in Higher Education Institutions of the Ukrainian Security Forces" [Gender Education of Cadets in Higher Education Institutions of the Ukrainian Security Forces]. While the focus of this analysis is on the Ukrainian security forces, it is pertinent to note that the analysis does not address the structural role of scientific research.

Following a thorough review of the abstracts, it was determined that the subsequent studies were to be excluded from consideration. While some of these studies were conducted within academic or institutional settings, it was observed that they failed to directly address the relationship between scientific research and institutional transformation within the military context. Additionally, these studies did not take into account the key variables that have been established in this review. In [19], "The Academic Performance of Cadets at the National Air Force Academy of the Republic of Angola," whilst situated in a military environment, focuses on the academic performance of cadets without analysing the role of scientific research as a driver of institutional change [19]. By[20] conducted an examination of approaches to material hardship classification.

Cross-sectional and longitudinal perspectives" addresses issues of social and economic classification, without any connection to military institutions or organisational transformation processes linked to research. In [21] conducted a study on the impact of Faculty Development Programmes (FDPs) on professional efficacy in developing countries. A comparative study among participants and non-participants of FDP in Bangladesh," focuses on faculty development programs in educational institutions in Bangladesh, without any connection to the armed forces or institutional change processes derived from scientific research. According to [22]. "A Librarian's Perspective on Sci-Hub's Impact on Users and the Library," undertakes an analysis of the impact of Sci-Hub on access to information. Notably, the analysis does not make any thematic connection to institutional transformation or the military environment.

By [23], "Preliminary investigation and analysis of the impact of the Sudan War on higher education and scientific research sectors", whilst mentioning scientific research, is situated within a context of conflict and general educational crisis, without exploring its structural role in processes of reform or institutional transformation in military environments [23]. Following a comprehensive review of the full text, the following studies were excluded from further consideration. While these sources initially appeared relevant due to their references to academic, scientific, or military contexts, it was determined that they did not meet the specific criteria of analytical depth, thematic relevance, and alignment with the central variables of this research. According to [24], "Analysis, classification, and philosophical foundations of critical thinking models", whilst providing valuable conceptual reflection on critical thinking, does not link its contributions to institutional development or the military or armed forces environment.

In [25], "Salvador Mazza, military doctor. Analysis of his professional career and scientific production in the Argentine Army Health Service," while focusing on a figure in the military sphere and his scientific output, takes a more historical and biographical approach than an analytical one regarding the impact of said output on contemporary institutional transformation processes. In the 2024 publication entitled "Educational Capital and International Mobility:", Khan provides a comprehensive analysis of the impact of educational capital on international mobility. A Bourdieusian inquiry into choosing peripheral higher education destinations," develops a sociological approach to international educational mobility from a Bourdieusian perspective, but does not consider the role of scientific research within military organisations or its institutional impact.

The selected studies were included in this systematic review because they directly address topics related to the academic and scientific environment in military contexts, such as training in military institutions, professional performance, educational innovation, organisational climate, the development of research capabilities, and strategic planning within the armed forces. All of the studies meet the established criteria for thematic and methodological relevance, as they explore key variables related to institutional transformation through scientific research. Furthermore, they present a solid methodological structure, offer relevant theoretical frameworks, and provide empirical contributions that strengthen the understanding of the role of academic research as a driver of change and modernisation within the Army and other armed institutions. In this regard, a table (Table 1) is presented with the main characteristics of the selected research.

Table 1: Characteristics of the studies analyzed according to the PRISMA methodology.

Authors and year	Title	Aim	Methodology	Results
Palma & Andrade (2020) [26]	Adaptação ao ensino superior militar: preditores do sucesso académico	To identify and characterize the stressors present in the effort to adapt to military higher education, analyzing how they evolve over the years and constitute possible predictors of academic success.	Quantitative, cross-sectional and applied approach, with both descriptive and explanatory objectives.	The longitudinal nature of the study emphasizes the results obtained and suggests that, in subsequent studies, correlations should be established with measures of self-esteem and performance as official measures in an operational context.
Shang (2021) [27]	Research on the Construction of Sci-tech Periodicals in Military Academies Based on Spearman Correlation Analysis of Academic Evaluation Index	Analyze the overall influence of scientific and technological (sci-tech) journals from military academies and how factors such as the proportion of funded articles affect their impact factor.	Quantitative approach Indicators such as total citation frequency, proportion of funded articles, external citation rate, and impact factor were collected and analyzed.	It was identified that the overall influence of these journals still needs improvement The proportion of funded articles has a positive effect on the impact factor.
Espitia Cubillos, Agudelo Calderon & Ramirez Contreras (2021) [28]	Perceptions of technological innovations in the Colombian Army	To detect expectations regarding innovative trends based on the perceptions of active members of the Army.	Approach: Mixed, combining interviews (qualitative) and surveys (quantitative). Scope of research: Exploratory and descriptive. Design: Cross-sectional experimental.	All the categories studied are important and cannot be eliminated or grouped into another or others, and those related to the defense of the community and the territory are of greatest interest to the participants of the study.
Marchisio & Spinello (2021) [29]	Internationalizatio n for Enhancing the European Security and Defence Higher Education	Identify which internationalization initiatives and policies can be developed and how they should be implemented to strengthen Higher Education in Security and Defense in Europe.	The experiences of the University of Turin over the past seven academic years were analyzed. Both quantitative and qualitative data were used.	The study highlights the positive impact of these initiatives on the development of academic networks and intercultural skills.
Reyes, Ariza, Alves (2021) [30]	Moral development in educational contexts: A systematic review	Identify the evidence available in the scientific literature related to the study of moral	Bibliographic database searches were conducted using predetermined search criteria,	The results suggest the influence of the educational context on moral development, from all levels of educational

		development in educational contexts.	following the PRISMA methodology.	training, encouraging the development of studies that enrich the scientific discussion.
Semenenko, Kirsanov & Chernyshova (2022) [31]	A Mathematical Model to Determine the Optimal Ratio of Researchers of Different Categories for Solving a Scientific Problem in the Military Sphere	To propose a variant of a mathematical model that justifies the optimal proportion of researchers from different categories to carry out scientific research of the highest possible quality, within a context of limited resources.	The methodology is based on the formulation of a mathematical model that describes the relationship between the quality of scientific research and the number of researchers, using a canonical parabola equation. The method of Lagrange multipliers is employed.	The results of the numerical experiment validate the usefulness of the model for improving research planning and performance in military institutions.
Surkov (2022) [32]	Elaboration of a method for strategic analysis of the development of the Armed Forces	Determine (clarify) the long-term strategy for developing the capabilities of the Armed Forces.	The methodology employs a conceptual approach that combines the selection of priority areas for capacity building with a recursively organized computational process.	The SADAF-RW method provides objective and quantifiable assessments of the effectiveness of strategic decisions, projecting the expected effects of the selected strategies.
Candumbo, Caimbo & Varona (2022) [33]	Pedagogical foundations of the professional military performance of the human resources of the Angolan Army Academy	To assess the pedagogical foundations of the professional military performance process of the human resources of the Military Academy of the Army (AMEX) of the Republic of Angola.	Using a qualitative methodology, documentary research is used to refer to the norms that support postgraduate education.	The pedagogical foundations of the advanced education theory regarding the professional military performance of human resources and its structuring in the practice of military education were identified.
Yahupov, Bielikov, Tkachenko (2023) [34]	Lifelong learning: models and methods of implementation	To present results of scientific research aimed at improving the efficiency of the educational process within the framework of continuing education, addressing key issues such as the development of organizational skills, adult education, foreign language learning, university professional development and receptiveness to change.	Various methodological approaches were used depending on the chapter. These included pedagogical experiments to analyze the development of organizational skills in military cadets, comparative analyses between the educational systems of Ukraine and Sweden, the design and implementation of innovative models such as the flipped classroom and methodologies based on technological immersion, as well as studies on university management and corporate training using active methods and coaching.	The studies made it possible to demonstrate progress in the development of key components of organizational competence in cadets, identify global and local trends in adult education, and propose effective models for language learning in multicultural contexts.
Calderón, Godoy, Marrero (2023) [35]	Validation of an Organizational Climate Scale in a Military Higher Education Institution: Implications for Research and Practice	Measuring the organizational climate in military higher education institutions in Ecuador.	The research is framed within the positivist paradigm. It employs a quantitative approach. The type of research is descriptive. The research design is based on a cross-sectional approach.	This study contributes to the understanding of the organizational climate in military higher education institutions in Ecuador and may have implications for the management and improvement of educational quality in the context of higher education in Ecuador.
Espiñeira, Muñoz & Garcia (2023) [36]	Perceptions and effectiveness of plagiarism detection mechanisms in Spanish, Portuguese and Ibero-American	To analyze the practices and perceptions regarding plagiarism detection in Ibero-American Social Sciences scientific journals indexed in Scopus, identifying the tools used, the editorial rejection	A mixed-methods approach was used.	The study revealed that most journals use plagiarism detection software, although there are doubts about its effectiveness.

	social science journals	criteria, and the effectiveness of current preventive measures.		
Kuchinski (2023) [37]	Exploring Perceptions on the Influences of Mentor-Mentee Relationships on Graduate-Student Learning	To explore graduate students' perceptions of how mentor-mentee relationships influence their learning within the U.S. Army's Master of Military Arts (MMAS) program.	A qualitative design of multiple case studies was used.	The study provides useful evidence to better understand the role of mentoring in academic settings, strengthen teacher development, and prepare mentors and students for their postgraduate academic journey.
Kuchinski (2024) [38]	A dinâmica da política de ensino: Uma perspectiva global da educação	To explore the dynamics of teaching politics in education at a global level, identifying approaches, challenges and good practices to assess its importance in the context of local political crises and global issues such as climate change, security and peace.	Narrative literature review.	The study provides an overview of the current state of political education in the world and recognizes the dynamics of its teaching.
Fernandez, Miron Sophía (2025) [39]	Contributions of the inductive method to the teaching of military ethics	Evaluate the impact of the Basic Concepts in Military Ethics course and its inductive approach on the attitudes and behaviors of Colombian military personnel.	A quantitative approach methodology was applied.	It was identified that the course had a positive impact on participants' attitudes regarding the importance of military ethics, its contribution to unit cohesion, and its influence on responsible decision-making.

Source: Authors, (2026).

In the domain of military higher education, numerous studies underscore the significance of enhancing academic adaptation, cultivating organisational skills, and fortifying pedagogical foundations. [26] identified stress factors that influence cadets' academic performance. In their 2021 study, [29] emphasised the role of academic internationalisation in the development of cooperation networks and intercultural skills in the European defence and security context. By [33] analysed the pedagogical foundations that underpin the training of military personnel in Angola. In addition, [34] have proposed innovative continuing education models with a view to improving the training efficiency of cadets and military professionals. In their 2023 study, Calderón et al. explored the organisational climate within military higher education institutions, offering practical methodologies to enhance educational standards.

As [37] demonstrated, through a series of case studies, the positive impact of mentorship on the learning of postgraduate students in military arts. By [39] demonstrated that the inductive approach to military ethics instruction fosters unit cohesion and responsible decision-making, thereby contributing to the moral development of military personnel. In the domain of scientific production and knowledge management within military institutions, [27] conducted a comprehensive analysis of the factors that influence the impact of military science and technology journals. The study revealed a direct correlation between the proportion of funded articles and their visibility. According to [36] addressed the issue of plagiarism detection in social science journals in Latin America, highlighting the widespread use of anti-plagiarism software, although with reservations about its effectiveness. In [31] developed a mathematical model to optimise the proportion of researchers based on the expected quality of the research, thus improving scientific planning in contexts with limited resources.

In the domain of technological modernisation and institutional perception, [28] investigated the perceptions of Colombian Army personnel regarding technological innovations, concluding that categories related to territorial defence are the most pertinent for the military. In [32] proposed a computational method (SADAF-RW) for the analysis and projection of the development of Armed Forces capabilities, providing a quantitative tool for long-term decision-making. In the domain of ethical, political and civic education, [30] conducted a systematic review that confirmed the influence of the educational environment on moral development at different levels of education. In the 2024 publication, Khanh (2024) explored the dynamics of teaching politics from a comparative perspective, emphasising its significance in the context of contemporary challenges, including security issues, climate change, and socio-political crises.

IV. DISCUSSION

IV.1 COMPARISON WITH OTHER STUDIES

The present research is distinctive in its adoption of a systemic and institutional perspective, a departure from the approaches of other academic works in this field. In contrast to [26], who address the factors of academic adaptation in cadets from an individual and cross-cutting approach, this work transcends the formative level to analyse how structured research processes impact organisational and doctrinal redesign. According to [39] emphasise the value of the inductive method in military ethics education, this research contends that such methods must be articulated within an institutional culture oriented towards critical thinking, within a cross-cutting axis of modernisation.

In relation to the research conducted by [29] on the internationalisation of defence education, the necessity for inclusive participation is emphasised, and structural reforms that integrate all institutional actors into open and technologically adaptive knowledge policies are proposed. According to [31], [32] have each contributed mathematical models for the optimisation of research and strategic planning. This work draws on these contributions to emphasise that their effectiveness is only guaranteed if they are integrated into a solid institutional science policy with robust normative and ethical structures. In the present study, [35], [34] analyse the impact of organisational climate and individual receptiveness to change. In contrast to the present study, which examines these psychosocial factors in relation to structural, political, and educational variables, this work posits that research should be positioned as a catalyst for organisational change rather than as an isolated effort.

IV.2 IMPLICATIONS OF THE STUDY

It is evident that scientific research, far from being a complementary function, must be institutionalised as a strategic capability that contributes to doctrinal development, evidence-based decision-making, and the professionalisation of military personnel. It is imperative that research policies are accompanied by structures that guarantee academic integrity, methodological validation, and the production of applicable knowledge. By [36] posit that the absence of rigorous control systems can compromise the credibility of scientific products, thereby diminishing their potential to influence real-world decisions. In this sense, the assertion by [33] regarding the need to ground professional development in research processes is fully relevant: advanced science-based training strengthens both operational capacity and institutional legitimacy.

IV.3 ADDED VALUE

In contrast to studies that concentrate on fragmented formative, normative, or technical aspects, this work integrates them to formulate a systemic vision of institutional change. A seminal contribution of this study lies in its establishment of a nexus between organisational culture and research transformation, thereby underscoring the pivotal role of critical thinking, academic mentorship (as articulated by [37]) and strategic political education (as expounded by [38]) as interwoven pillars of military leadership. Moreover, it puts forward an interpretation that is firmly embedded within the Peruvian context, yet with a pronounced emphasis on the opportunities presented by existing human capital, as articulated by [28].

IV.4 RECOMMENDATIONS FOR FUTURE STUDIES

It is recommended that further longitudinal studies be conducted in order to evaluate the real impact of scientific research on operational indicators of the Army. Such indicators may be categorised as follows: logistical effectiveness, deployment capacity, and doctrinal innovation. In the context of Peruvian armed forces, the undertaking of comparative studies among Latin American armed forces with varying degrees of scientific institutionalisation would prove pertinent in order to extract models that are adaptable to the Peruvian context. Another under-explored area is the evaluation of incentive and recognition policies for research within military careers, as well as their impact on the motivation and retention of scientific talent. It is recommended that efforts be undertaken to develop return-on-investment models for military research that evaluate not only the technical impact but also the effect on organisational cohesion, institutional legitimacy, and civil-military cooperation.

V. CONCLUSIONS

Scientific research contributes significantly to institutional transformation in the Peruvian Army by strengthening the academic and professional training of military and civilian personnel through the implementation of strategically planned educational processes that foster critical thinking, ethical reflection, and the generation of relevant knowledge. This process has been shown to enhance curricular content, whilst concomitantly effecting a transformation in the axiological and leadership profile of future officers. This, in turn, serves to promote an organisational culture oriented toward continuous improvement. Institutional transformation driven by research also requires the coordinated participation of multiple actors within the organisation and the adoption of inclusive policies that respond to new realities. This is necessary to guarantee the comprehensive, ethical, and sustainable development of institutional capabilities.

The findings of scientific research have been shown to exert a considerable influence on the processes of modernisation and decision-making in the Peruvian Army. This influence is evidenced by the provision of methodological and quantitative tools that serve to optimise the allocation of resources, enhance strategic planning, and fortify the quality of the knowledge generated. Research facilitates an objective evaluation of institutional capabilities, fosters transparency and effectiveness in decision-making, and contributes to enhancing academic prestige through high-impact publications. Moreover, the scientific and pedagogical training of military personnel fosters an organisational culture predicated on evidence, continuous improvement, and innovation – pivotal aspects for sustainable modernisation and decision-making consistent with contemporary challenges.

The Peruvian Army members' perception of the impact of scientific research on institutional transformation is influenced by a number of factors. These include their level of receptiveness to innovation, the organisational climate, and their training experiences in academic or mentoring contexts. While certain sectors are proactive in their promotion of change by acknowledging the strategic significance of science in the modernisation of procedures, technologies and doctrines, others adopt a more conservative approach due to hierarchical rigidity or structural limitations that impede participation. However, there is a general perception of the Army's innovative potential, based on its skilled and committed human capital and the need to integrate scientific research not only into technical aspects but also into the political and strategic training of personnel. This would promote a comprehensive, critical and adapted institutional transformation to contemporary challenges.

Financial support

The research was conducted without financial support from any institution; the researchers themselves covered all necessary expenses throughout the research process.

Conflicts of interest

The research was conducted without any conflicts of interest that could have potentially influenced the results.

Data availability

The data are available to any researcher who wishes to use them for methodological and scientific purposes

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Michael Joaquin Herrera Contreras, Aaron Aguirre Venturi and Ada Gallegos.
Methodology: Michael Joaquin Herrera Contreras and Aaron Aguirre Venturi.
Investigation: Michael Joaquin Herrera Contreras and Aaron Aguirre Venturi.
Discussion of results: Michael Joaquin Herrera Contreras, Aaron Aguirre Venturi and Ada Gallegos.
Writing – Original Draft: Michael Joaquin Herrera Contreras.
Writing – Review and Editing: Michael Joaquin Herrera Contreras and Aaron Aguirre Venturi.
Resources: Aaron Aguirre Venturi.
Supervision: Aaron Aguirre Venturi and Ada Gallegos.
Approval of the final text: Michael Joaquin Herrera Contreras, Aaron Aguirre Venturi and Ada Gallegos.

VII. REFERENCES

[1] L. S. , Z. C. J. A. , A. restrepo, Y. , J. N. B. , & P. S. E. G. Atencio Ortiz, "Revisión de mecanismos de gestión de ciencia, tecnología e innovación en el sector defensa," *Revista Logos, Ciencia & Tecnología*, vol. 17, no. 1, pp. 103–118, 2025, doi: 10.22335/rlct.v17i1.2017.

[2] F. Huayán, "El Desarrollo Tecnológico en el Ejército del Perú: Una Perspectiva Estratégica para la Transformación Institucional," *Revista Seguridad y Poder Terrestre*, vol. 2, no. 3, pp. 21–36, 2023, doi: 10.56221/spt.v2i3.32.

[3] G.-I. Toroi, "Methodological Approaches in Military Science Research. BULLETIN OF "CAROL I," *NATIONAL DEFENCE UNIVERSITY*, vol. 13, no. 3, pp. 89–104, 2024, doi: 10.53477/2284-9378-24-32.

[4] A. L. Pereira and Y. S. D. Sousa, "The Potential For Dual Use of Emerging Technological Innovations in The Brazilian Army," *Revista de Gestão Social e Ambiental*, vol. 19, no. 1, 2025, doi: 10.24857/rgsa.v19n1-160.

[5] J. Florez, L. Molina, and K. Gallego, "Towards a competitive ST system in the defense sector. Air Force case study," *Ingeniería Solidaria*, vol. 17, no. 2, pp. 1–28, 2021, doi: 10.16925/2357-6014.2021.02.06.

[6] R. Alonso, "Cultura Organizacional: Elemento Esencial en el Proceso de Transformación del Ejército del Perú," *Revista Seguridad y Poder Terrestre*, vol. 2, no. 4, 2023, doi: 10.56221/spt.v2i4.44.
[7] D. Educación and D. Ejército, *Directiva General de Funcionamiento del Sistema Educativo del Ejército*. Ejército del Perú, 2024.

[8] N. B. Proaño Rodriguez and J. F. Alvarez Valenzuela, "LA SEGURIDAD Y DEFENSA, EL DESARROLLO TECNOLÓGICO, LA INNOVACIÓN Y LA TRANSFORMACIÓN," *Revista de La Academia Del Guerra Del Ejército Ecuatoriano*, vol. 16, no. 1, 2023, doi: 10.24133/AGE.VOL16.N01.2023.03.

[9] M. Trujillo and G. Valencia, "Cambiando de mentalidad. El proceso de transformación del Ejército Nacional de Colombia (2011-2018," *Revista Científica General José María Córdova*, vol. 19, no. 36, pp. 867–887, 2021, doi: 10.21830/19006586.804.

[10] B. Jiménez, E. Villa, and J. Bermúdez, "La gestión de la tecnología y la innovación en el sector defensa: resultados desde un análisis bibliométrico," *Revista Virtual Universidad Católica Del Norte*, vol. 59, pp. 45–70, 2020, doi: 10.35575/rvucn.n59a4.

[11] J. Calizaya, "Algunas ideas de investigación científica," *Minerva*, vol. 1, no. 3, pp. 35–39, 2020, doi: 10.47460/minerva.v1i3.15.

[12] B. Zavala, B. Soledispa, P. Soledispa, and G. Sarmiento, "Educational management as a driver of scientific and technological research: New horizons for innovation," *Salud, Ciencia y Tecnología*, vol. 2, 2022, doi: 10.56294/saludcyt2022165.

[13] M. J. Page, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *International journal of surgery*, vol. 88, p. 105906, 2021, doi: 10.1016/j.ijsu.2021.105906.

[14] X. Levon, "Improving Language Learning with Virtual Reality: A Quasi-experimental Study," 2023.

[15] I. P. Bolaños Ramírez, A. F. Jiménez Vélez, and M. F. Noboa González, "Fuerza Aérea Ecuatoriana: en camino al multidominio. Un análisis transdisciplinario en complejidad," *Ciencia y Poder Aéreo*, vol. 17, no. 2, 2022, doi: 10.18667/cienciaypoderaereo.755.

[16] P. C. C. Castillo, "Implementación de capacitaciones en seguridad y salud en el trabajo para mejorar la cultura de gestión de seguridad del personal de la unidad de montaje de cableado de la empresa A\&{N} Proyectos S.A.C Talara," 2019.

[17] M. A. Mohaimen, I. Mahmud, M. S. Hassan, M. B. Amin, and S. K. Roy, "Unveiling HRM dynamics through PLS model: The moderating roles of self-efficacy and organizational commitment in work engagement and talent turnover intention," *Social Sciences & Humanities Open*, vol. 11, 2025, doi: 10.1016/j.ssaho.2025.101327.

[18] V. L. Kozachuk, A. L. Tsybizov, V. F. Tarhonskyi, V. P. Didichenko, V. M. Lushnichenko, and U. B. Dobrovolskyi, "Gender Education of Cadets in Higher Education Institutions of the Ukrainian Security Forces," *Health Education and Health Promotion*, vol. 10, no. 1, pp. 161–166, 2022.

[19] A. Fernando, "El rendimiento académico de los cadetes de la Academia de la Fuerza Aérea Nacional de la República de Angola," *Varona. Revista Científico Metodológica*, vol. 70, pp. 67–70, 2020.

[20] M. Grant, K. Meissel, and D. Exeter, "Examining approaches to material hardship classification: Cross-sectional and longitudinal perspectives," *Social Sciences & Humanities Open*, vol. 10, 2024, doi: 10.1016/j.ssaho.2024.100931.

[21] M. H. A. Rahman, "Faculty development programs (FDP) in developing professional efficacy: A comparative study among participants and non-participants of FDP in Bangladesh," *Social Sciences & Humanities Open*, vol. 7, no. 1, 2023, doi: 10.1016/j.ssaho.2023.100499.

[22] K. Okere and G. Evans, "A Librarian's Perspective on Sci-Hub's Impact on Users and the Library," *Ser Libr*, vol. 81, no. 1, pp. 112–117, 2021, doi: 10.1080/0361526X.2021.1963907.

[23] H. E. E. Abugabr Elhag and R. M. H. Baleela, "Preliminary investigation and analysis of the impact of the Sudan War on higher education and scientific research sectors," *Int J Educ Dev*, vol. 112, 2025, doi: 10.1016/j.ijedudev.2024.103190.

[24] A. M. Rodríguez-Ortiz, J. C. Hernández-Rodríguez, A. M. López-Rúa, and V. Cadavid-Alzate, "Análisis, clasificación y fundamentos filosóficos de los modelos de pensamiento crítico," *Sophía*, vol. 35, pp. 211–248, 2023, doi: 10.17163/soph.n35.2023.07.

[25] G. Soprano, "Salvador Mazza, médico militar. Análisis de su carrera profesional y producción científica en el servicio de sanidad del ejército argentino," *Folia Histórica Del Nordeste*, vol. 48, pp. 69–93, 2023, doi: 10.30972/fhn.48487079.

[26] A. Palma Rosinha and H. Andrade Júnior, "Adaptação ao ensino superior militar: preditores do sucesso académico," *Innovar*, vol. 30, no. 76, pp. 131–142, 2020, doi: 10.15446/innovar.v30n76.85218.

[27] F. Shang, "Research on the Construction of Sci-tech Periodicals in Military Academies Based on Spearman Correlation Analysis of Academic Evaluation Index," in *2021 2nd International Conference on Education, Knowledge and Information Management (ICEKIM*, 2021. doi: 10.1109/ICEKIM52309.2021.00155.

[28] A. A. Espitia Cubillos, J. A. Agudelo Calderón, and T. Ramírez Contreras, "Percepciones sobre innovaciones tecnológicas en el Ejército colombiano," *Revista Logos, Ciencia & Tecnología*, vol. 13, no. 2, pp. 85–102, 2021, doi: 10.22335/rlct.v13i2.1408.

[29] M. Marchisio and E. Spinello, "Internationalization for Enhancing the European Security and Defence Higher Education," in *International Conference E-Learning*, 2021, pp. 99–106.

[30] L. Lizeth Reyes-Ruiz *et al.*, "Desarrollo moral en contextos educativos: Una revisión sistemática," *European Journal of Child Development, Education and Psychopathology*, pp. 1–32, 2021, doi: 10.32457/ejpad.v9i1.1508.

[31] L. Semenenko, S. Kirsanov, A. Solomitsky, P. Onofriichuk, and I. Chernyshova, "A Mathematical Model to Determine the Optimal Ratio of Researchers of Different Categories for Solving a Scientific Problem in the Military Sphere," *Journal of Curriculum and Teaching*, vol. 11, no. 3, 2022, doi: 10.5430/jct.v11n3p103.

[32] O. Surkov, "Elaboration of a method for strategic analysis of the development of the Armed Forces," *Novum Jus*, vol. 16, no. 3, pp. 75–105, 2022, doi: 10.14718/NovumJus.2022.16.3.4.

[33] J. Candumbo, G. Caimbo, and A. Caro, "Fundamentos pedagógicos del desempeño profesional militar de los recursos humanos de la Academia del Ejército de Angola," *VARONA*, vol. 4, pp. 1–11, 2022.

[34] V. Yahupov *et al.*, *Lifelong learning: models and methods of implementation*. PRIVAT COMPANY TECHNOLOGY CENTER, 2023. doi: 10.15587/978-617-7319-70-1.

[35] D. Calderón, M. Godoy, and A. Marrero, "Validation of an Organizational Climate Scale in a Military Higher Education Institution: Implications for Research and Practice," in *CHAKIÑAN, REVISTA DE CIENCIAS SOCIALES Y HUMANIDADES*, vol. 22, 2023, pp. 128–144. doi: 10.37135/chk.002.22.08.

[36] E. M. Espiñeira Bellón, J. M. Muñoz Cantero, A. M. Porto Castro, and M. J. Mosteiro García, "Percepciones y eficacia de los mecanismos de detección de plagio en revistas científicas de Ciencias Sociales españolas, portuguesas e iberoamericanas," *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa*, vol. 29, no. 2, 2023, doi: 10.30827/relieve.v29i2.29097.

[37] W. Kuchinski, *Exploring Perceptions on the Influences of Mentor-Mentee Relationships on Graduate-Student Learning*. Kansas State University, 2023.

[38] L. Khanh, "A dinâmica da política de ensino: Uma perspectiva global da educação," *Synesis*, vol. 16, no. 2, pp. 62–89, 2024.

[39] A. E. Fernández-Osorio, M. Miron, and D. Whetham, "Aportes del método inductivo a la enseñanza de la ética militar," *Sophía*, vol. 38, pp. 79–105, 2025, doi: 10.17163/soph.n38.2025.02.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 856-868. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3095

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

EARLY ALZHEIMER'S DISEASE DETECTION BY USING RESNET FOR EXTRACTION OF FEATURES WITH SVM TUNING FOR CLASSIFICATION

Rama Lakshmi B[*1], Ch Rama Devi[2], Malathi Janapati[3], Mirtipati Satish Kumar[4] Anjani Y[5], Pavan Gunda[6]

[1, 2, 5, 6]Department of Computer Science and Engineering, Sir C R Reddy College of Engineering, Eluru-534007, AP, India
[3]Department of Computer Science and Engineering, Siddhartha Academy of Higher Education, Deemed To Be University, Vijayawada, India
[4]Department of Information Engineering and Computational Technology, MVGR College of Engineering, Vizianagaram-535005, AP, India

[1]https://orcid.org/0000-0001-7041-2260, [2]https://orcid.org/0000-0003-4476-4363, [3]https://orcid.org/0000-0001-7096-0435
[4]https://orcid.org/0009-0000-2920-7812, [5]https://orcid.org/0000-0003-1006-4051, [6]https://orcid.org/0009-0008-6163-6789

Email: *ramalakshmi.boyapati@gmail.com, ramadevi21vg@gmail.com, malathi.j@vrsiddhartha.ac.in, satishkumarsrk@icloud.com, anjaniyalamanchili@gmail.com, gsraghupavan9@gmail.com

ARTICLE INFO

Article History
Received: December 10, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Tuning,
Resnet,
SVM,
Neuro Imaging,
adaptive Filters,
MRI,
Hyperopt.

ABSTRACT

Early detection of Alzheimer's disease (AD) is vital for effective intervention and management. This study introduces a novel approach that employs a ResNet architecture for feature extraction from neuroimaging data, coupled with Support Vector Machine (SVM) tuning for classification. We utilized a dataset of MRI scans from subjects at various stages of cognitive decline. The ResNet model was fine-tuned to identify complex patterns in brain structure indicative of early AD. Features extracted from the network's final layers were then input into an SVM classifier, optimized through grid search to improve classification accuracy. The parameters are tuned and the final layer in Resnet-53 is used for Support Vector Machine (SVM) model. It is mainly used for detection of Alzheimer's, and it addresses the both binary and multiclass classification. But we are using multiclass labels are used so it is used for multiclass classifications. Bayesian optimization with Hyperopt is used in a fine tune process to find the hyperparametrs space, optimizing key variables such as kernel selection and regularization to increase the performance of the model on the validation set. The proposed method is used to show the most accurate difference between normal cognitive aging and Alzheimer's disease, which is so sensitive and specific to achieve. Multimodal data is combined to increase the models performance and provides comprehensive tools for early detection.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Alzheimer's disease (AD) is a increasing neurodegenerative condition described by intellectual regression, memory loss, and weakened decision-making. It is one of the heading causes of dementia, involving millions of people worldwide [1]. The prompt and correct diagnosis is important for appropriate interference and advanced patient outcomes. However, the conventional diagnostic methods, involving clinical assessments and neuropsychological tests, are frequently biased and may lack correctness. The advanced machine learning (ML) techniques, specifically deep learning models, have developed as effective tools for improving the accuracy and efficiency of AD detection [2]. In current years, convolutional neural networks (CNNs) have been extensively used for medical image analysis, exhibiting excellent performance in feature extraction and classification tasks. Among these, ResNet-50, a deep CNN architecture, has shown outstanding efficiency in capturing complicated patterns from neuroimaging data, predominantly from MRI scans [3].

In this study, ResNet-50 is utilized for feature extraction, using its deep categorized structure to distinguish between different stages of Alzheimer's disease. Besides, Support Vector Machine (SVM), a robust classification algorithm, is enhanced using Bayesian hyperparameter tuning (Hyperopt) to improve classification performance. This research aims to develop a reliable ML-based framework for automated AD detection, grouping patients into four groups: Non-demented, Very Mild Demented, Mild Demented, and Moderate Demented. By incorporating deep learning with optimized classification techniques, the proposed approach improved the sensitivity and specificity, by this means assisting clinical decision-making and progressing diagnostic methodologies in neurodegenerative diseases.

I.1 AD DISEASE DETECTION USING TRADITIONAL CNN:

It describes a process in which MRI samples are converted into JPEG slices using MATLAB, rescaling to 8-bit pixel size from the original 14-bit size. The three blocks that comprise the proposed model are an ensemble of convolutional, ReLU activation, & max pooling layers in each block. A flattening conv layer, two fc layers, two sets of convolutional and max pooling layers, and a softmax/sigmoid classifier are all included in the design. MCI, NC, & AD are the output classifications [4]. The input is a 256x256 grayscale image that undergoes conv layers with 32 mappings for features. The filter size is 3x3, and ReLU activation. Max pooling is applied with a size of filter 3x3, resulting in image dimensions of 84x84x32. Subsequent layers include a 2nd conv layer with 64 FM & a 3rd convolutional layer with 128 FM, connected to max pooling [5]. The flattened layer has 8192 parameters. A conv layer containing 256 1x1 feature maps is used as the fifth layer, which is fully coupled to the layer before it. The last layer is a completely linked softmax output layer of 10 potential values, while the sixth surface is an entirely linked layer with 256 units. The overall architecture for this methodology is presented in figure 1.

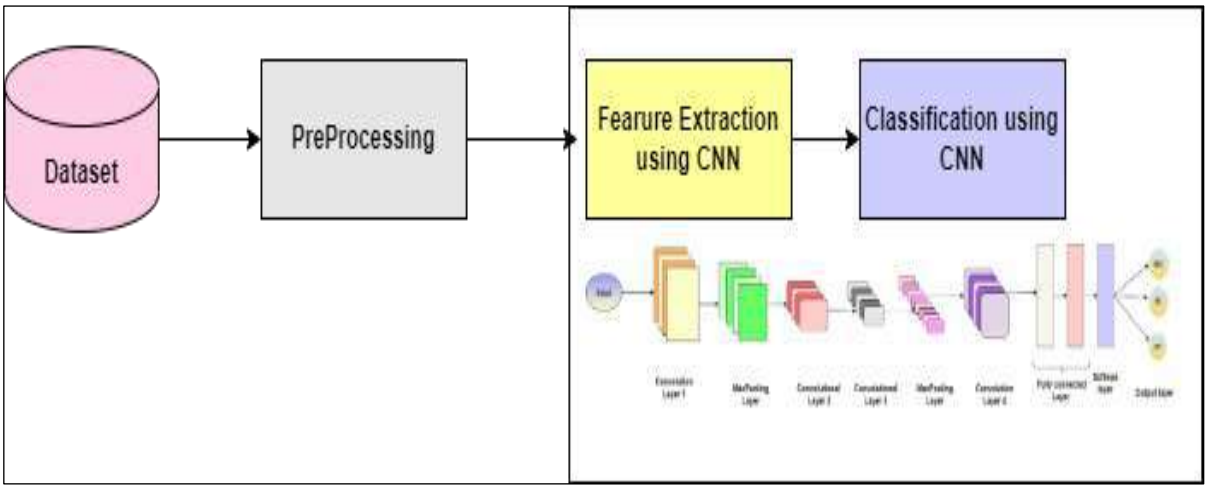


Figure 1: Traditional CNN along with Feature Extraction.
Source: Authors, (2026).

I.2 AD DISEASE DETECTION USING INTEGRATED CNN WITH CAD:

A CAD (Computer-Aided Diagnosis) device utilizing a convolutional neural network (CNN) was developed to distinguish Alzheimer's Disease (AD) from normal control (NC) patients. The CAD system analyzed 18FDG-PET images and employed a CNN to extract relevant features. The pictures were divided into several 2D slices and clustered at regular intervals without crossing. The ADNI dataset was used to assess the proposed CAD system. The suggested CNN-based CAD system performed exceptionally well in simulations, outperforming other techniques in terms of precision in classification and resilience [6]. The findings strongly indicated that the proposed system had a higher potential for effectively differentiating AD from NC patients.

I.3 ALZHEIMER'S DISEASE DETECTION USING CUSTOMIZED CNN:

Data pre-processing was performed on the OASIS dataset to address the influence of different image sizes. [3] Image resizing using OpenCV Python was employed to reduce training time, while image denoising was conducted to improve model performance. The dataset was labelled for binary classification based on CDR 0 | 1, representing Non-Demented and severe Alzheimer's cases, respectively. An equal number of patients with Alzheimer's Disease and those without dementia were included in the training and testing datasets, which were divided into 8:2 ratios and developed a 12-layer CNN structure to classify and identify Alzheimer's Disease in binary form. [2]

Model used data pre-processing methods such as picture scaling and image denoising for our investigation, which used the OASIS dataset [7]. The suggested model combines machine learning and deep learning techniques. It performs better than four previously trained CNN models and a pre-existing 8-layer CNN model. Model intend to broaden our investigation in future work and concentrate on the untimely diagnosis of AD using multiple classification classes using the OASIS dataset. Figure 2 presents the Customized CNN architecture.

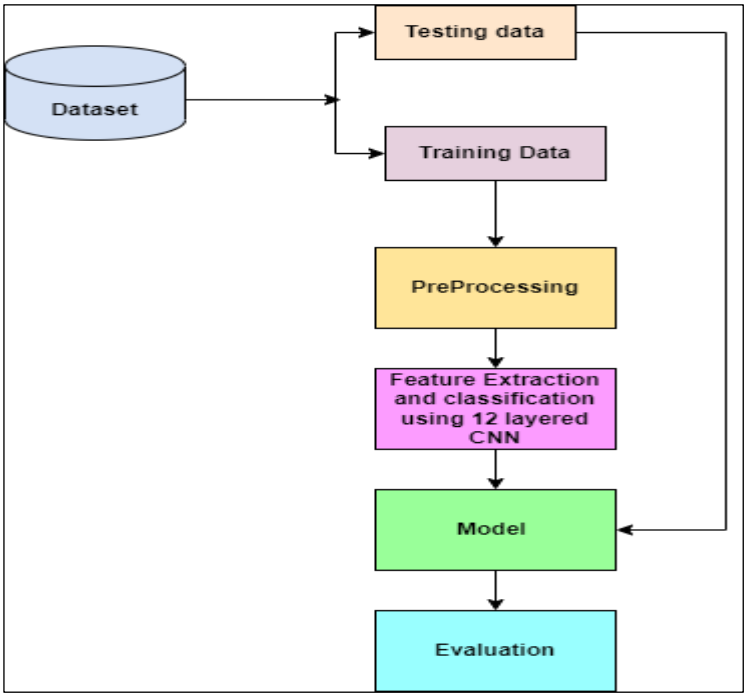


Figure 2:12-layered CNN for Disease Detection.
Source: Authors, (2026).

I.4 ALZHEIMER'S DISEASE DETECTION USING HYBRID MODELS:

A hybrid signal methodology called HEMRDTL is proposed for the analysis & classification of AD. [8] The model combines EEG-Fused CT/MRI data and utilizes transfer learning with a VGG-19 TranL deep neural network. The HEMRDTL model outperforms other transfer learning models in terms of accuracy. Future research using genetic biomarkers & OCT images for early detection of AD is mentioned, highlighting the differentiation of the eye region in individuals with AD and the relevance of genetic biomarkers for diagnosis and prognosis [9]. Obtaining a series of OCT images showing people with AD is a goal for further research in this important area. Leveraging additional data sources could enhance the precision & viability of early AD detection. Figure 3 presents the Hybrid model detection

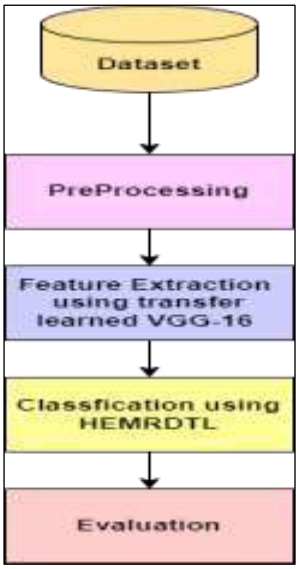


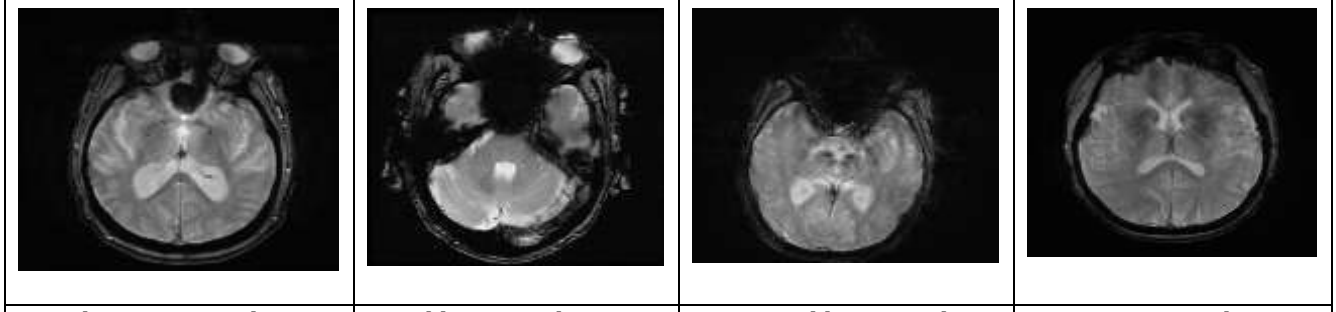
Figure 3: VGG-16 for Feature Extraction + HERMDTL for Classification.
Source: Authors, (2026).

One of the newest approaches to managing high-dimensional information and distinguishing between healthy people and AD patients is using machine learning (ML) techniques. The benefit of the method known as machine learning is the effectiveness of a proficient system: given adequate data, the algorithm can be trained to learn a tremendous amount of knowledge in a relatively short period (the learning process usually may take a few minutes to a few days at most); in contrast, to train an experienced neurologist/clinician frequently requires years of hard work and internships. The programme backed by machine learning may significantly lower the cost of diagnostics [10]. The decision-making phase of such classification or recognition only takes a few seconds, which is now one of the most alluring advantages over a traditional doctor. A typical ML technique typically explores features to discover the most critical departing patterns between MCI/AD and normal control subjects.

DATASETS AVAILABLE:

a. AIBL (Australian Imaging and Biomarker Lifestyle): The AIBL PiB and MRI images were captured using ADNI methods, but no post-processing was used. With the help of funding from the Alzheimer's Association, this data publication has become possible. The AIBL dataset utilized in this investigation included 37 participants' 3T T1-weighted MRI scan results, which included 25 sMCI and 12 pMCI. The study collects a wide range of participant data, including demographic information, medical history, cognitive assessments, blood samples, genetic information, brain imaging (MRI and PET scans), and cerebrospinal fluid (CSF) samples.

b. ADNI (Alzheimer's Disease Neuroimaging Initiative): The dataset utilized in this investigation was given by the ADNI (AD Neuroimaging Initiative) study. The most often used dataset for research in this area is ADNI. Its major objective is to ascertain if it is feasible to monitor AD progression by combining neuropsychological, clinical, and other tests with MRI, PET, and other biological markers. [1] The public-private collaboration-based investigations ADNI1, ADNI-GO, and ADNI2 are continued in ADNI3, which was previously sponsored. The ADNI study collects a wide range of data from participants, including clinical and demographic information, neuropsychological assessments, medical history, genetic data, blood samples, cerebrospinal fluid (CSF) samples, structural and functional MRI scans, and positron emission tomography (PET) scans. The dataset also includes data from various biomarker assays, such as amyloid beta and tau protein measurements. Figure 4 represents different classes available in the OIASIS Dataset. And Table 1 shows the information about different datasets available for the Alzheimer's.



Moderate Demented	Mild Demented	Very Mild Demented	Non Demented

Figure 4: Different Classes Available in the OIASIS Dataset.
Source: Authors, (2026).

Table 1: Dataset Information for Alzheimer's.

S. NO	Dataset	No. of attributes or images	Class label
1.	AIBL	1335	AD, CN
2.	ADNI	94	LMCI, MCI, EMCI, AD, CN
3.	OIASIS	140,000	LMCI, SMC, EMCI, AD, CN
4.	MIRIAD	708	AD, CN

Source: Authors, (2026).

II. LITERATURE SURVEY

Vijeeta Patil et al. [11] have focused on different transfer learning methods and applied them to deep learning. The data was collected from Kaggle and ADNI, where 6k samples are present in the dataset. Here the data contains four different modules while extracting the data will be examined by combining the whole module. The data is initiated by inputting the data to the feature extraction mode. At the initial stage, the data need to be mapped then the pooling of features is generated. The process is continued, and finally, the data is flattering for classification. The data is fully connected in the next stage, and the classification is performed based on weights. Finally, the data is in a probabilistic mode, where the weights are generated based on the types. Identifying neuron diseases at every stage is an essential task for everyone.

Alzheimer's Disease is tuff and difficult to cure, so it needs immediate attention for identification. By [12] HEMRDL model works with two acquisition and transfer learning phases called VGG19. The data was collected from OASIS, which contains both diseased and healthy person MRI. In data acquisition, the model is generated based on images, and analysis is made and created in a graph format. Then the data is inputted into a pre-processing stage, where it removes all unwanted data in two forms image and graph. Parallelly, VGG19 is applied to images, and based on threshold values, new patterns are generated. The data is now ready for feature extraction, combining transfer learning and pre-processing. The extracted features of images and graphs are connected to classification, where the fc is applied in three stages, and the four different phases are retrieved.

Finally, the data is evaluated based on the reports and doctors' analysis. The brain is the main part of controlling the whole body with the help of neurons. Alzheimer's Disease is related to the brain and can cause serious attacks if not cured at the initial stage. [13] have proposed a transfer learning method used in the Convolution model where the Disease can be identified initially. The dataset contains 1k images related to MRI from ADNI. It contains more than four categories. The data were pre-processed to remove unwanted data, and MRI processing to 2d slices. The data is now a gm slice, classified based on the deep neural networks with transfer learning. [14] The CNN contains 15 layers where parallelly the transfer learning denseness is used.

Finally, the data is predicted based on the four categories. Hence the performance is increased based on the image segmentation with good accuracy.

Table 2: Pros and Cons Analysis on different models.

Author	Algorithm	Merits	Demerits	Accuracy
M Leela et al.	VGG19, HEMRDTL	By evaluating separately, the data has acquired high performance.	By leveraging the data, the accuracy can be increased.	99%
Sheng Liu et al	DL	Images can be automatically detected.	Performances has to be improved.	83%
Ahila A et al	CNN	This model can solve several issues at a time.	The dataset was not related to humans.	96%
P C Muhammed Raees et al	SVM and DNN	Implemented to any application.	Based on the datasets the performance is being different.	80%
Ahmad Salehi et al.	CNN	Combines different datsets and performance is increased.	Execution time is more because of huge datsets.	97%
Noman Raza et al.	DenseNet	By initialling the epochs, the performance can be easily estimated.	The test has been performed in the final stage, which may decrease the accuracy.	97%
M. Ghazal et al	Alex Net	Handcrafted features are not used.	Complexity of time is more because of large images.	98%

Source: Authors, (2026).

Objectives:

This study aims to develop an ML-based approach for AD detection using ResNet-50 for feature extraction and Support Vector Machine (SVM) for classification. The objective is to enhance accuracy, sensitivity, and specificity in distinguishing between Non-demented, Very Mild Demented, Mild Demented, and Moderate Demented cases.

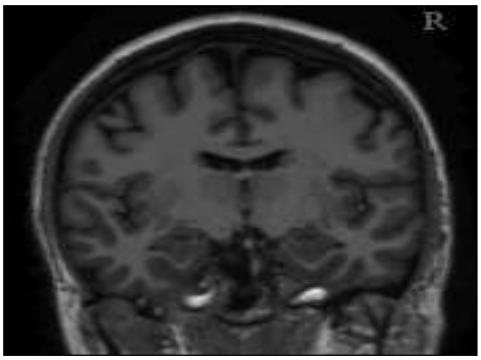
- **Objective 1:** To implement tuned pre-trained model, ResNet for performing feature extraction on OASIS data to segment the diseased cells
- **Objective 2:** To develop Enhanced Hyperopt based transfer learning model for classification of disease. Enhanced Hyperopt technique helps the model to control the parameters of neural networks based on multi objective
- **Objective 3:** To comparative analysis of existing Approaches with Proposed Approaches of Alzheimer Disease detection.

III. PROPOSED METHODOLOGY

System architecture

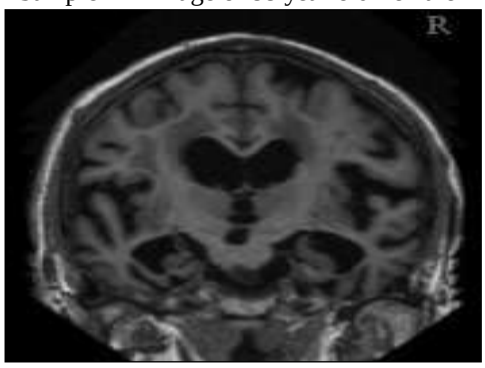
OASIS dataset which is an open access database. OASIS data contains Alzheimer related features along with the different gender wise disease prediction. Open Access Series of Imaging Studies (OASIS). The OASIS project makes available for research the datasets of brain MRI of normal and demented patients. [15]The goal is to communicate new neuro scientific discoveries and developments. The Washington University Alzheimer Disease Research Center collaborated to create this open access database. System architecture is clearly shown in the figure 6.

Sample AD image of 21year age Male



CDR=0 (Normal case)

Sample AD image of 88 year old Female



CDR=0.672 (mild case)

Figure: 5 Sample Alzheimer Disease of different age groups.
Source: Authors, (2026).

Price discounts and scores. The similarities make the new device more accurate, versatile and compliant with present day e-commerce requirements, together with solving the restrictions of the old gadget greater effectively.

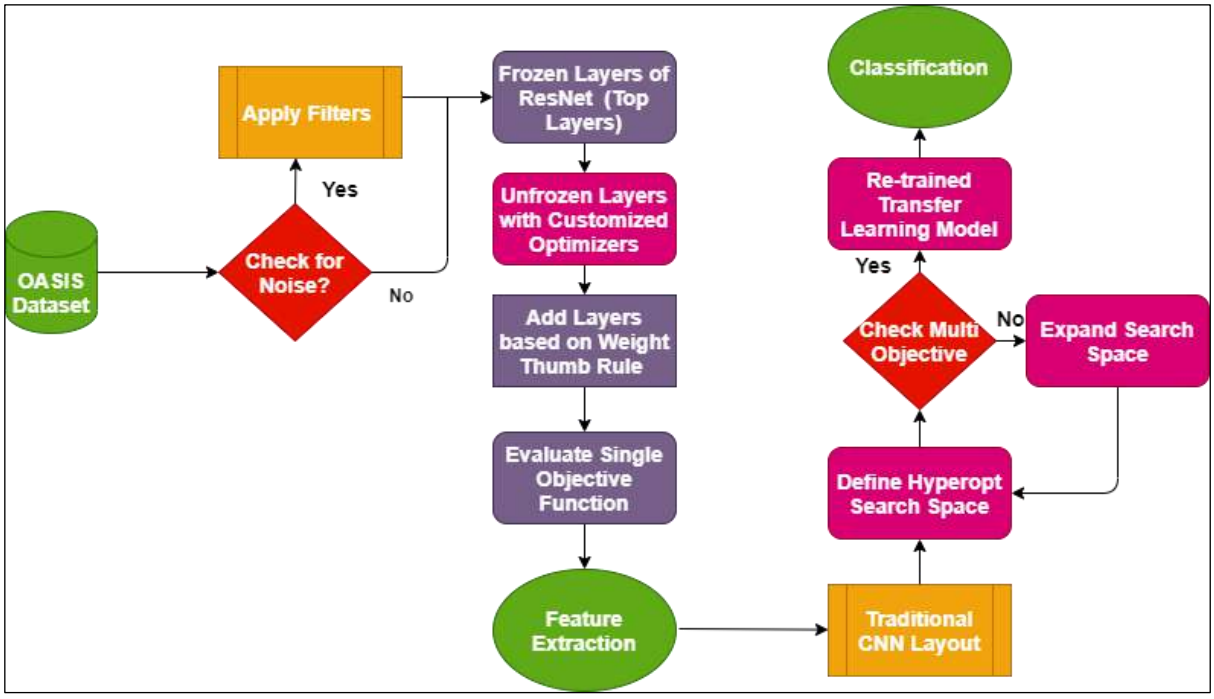


Figure 6: System Architecture.
Source: Authors, (2026).

For feature extraction and classification, a modified ResNet-50 architecture is employed as shown in Figure 6 [16]. The top layers of ResNet-50 are frozen to retain pre-trained weights, while three additional layers are added. Among these, two layers focus on feature extraction, and the final layer is dedicated to classification. To enhance model optimization, a hyperparameter search space is implemented using hyperparameter tuning techniques [17]. Traditional machine learning models such as Support Vector Machines (SVMs), neural networks, Bayesian models, expectation-maximization algorithms, and decision trees have been widely used for feature selection and disease classification [18]. However, as the number of noisy or irrelevant features increases, these models face challenges in computational efficiency and pattern evaluation.

Preprocessing and Model Development

The system architecture, as depicted in Figure 6, is structured to enhance image quality and optimize model training. [19] The dataset used in this study, titled "Alzheimer_s Dataset-20231129T135615Z-001," undergoes an initial preprocessing stage where noise is reduced using a combination of bilateral and Gaussian filters, followed by zero-centring normalization. The Gaussian filter is mathematically expressed as:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\left(\frac{x^2+y^2}{2\sigma^2}\right)}$$

Where:

- $G(x,y)$ represents the Gaussian kernel at position (x,y).
- σ ss the standard deviation that controls the amount of smoothing.
- $e^{-\left(\frac{x^2+y^2}{2\sigma^2}\right)}$ determines the weight based on the pixel distance.

To apply Gaussian filtering, the convolution operation is performed:

$$I_{filtered}(x,y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} I(x+i, y+j) G(i,j)$$

Where:

- $I_{filtered}(x,y)$ is the filtred intensity pixel (x,y)
- $I(x+i, y+j)$ represents neighboring pixel intensities.
- $G(i,j)$ is the Gaussian Kernel weight

Bilateral filter:

The bilateral filter can be mathematically expressed as:

$$I_{filtered}(x) = \frac{1}{W_p}\sum_{i\in\Omega} I(i) f_s(||x-i||) f_r(|I(x)-I(i)|)$$

Where:

- $I_{filtered}(x)$ is the filtered intensity at pixel x
- $I(i)$ is the intensity of the neighboring pixel i
- $f_s(||x-i||)$ is the spatial weight based on the Euclidean distance between pixels.
- $f_r(|I(x)-I(i)|)$ is the range weight based on the intensity difference.
- W_p is the normalization factor:

$$W_p = \sum_{i\in\Omega} f_s(||x-i||) f_r(|I(x)-I(i)|)$$

Oasis Data as (Alzheimer_s Dataset-20231129T135615Z-001) is taken. Initially Noise is reduced by using two filters combination as bilateral and Gaussian filters and used Zero Centering. Then the pre-process will be done by using Resnet-50 by adding three layers to it, top layers are frozen to calculate the weights and unfrozen layers are customized. In added three layers 2 layers are used for feature extraction and one layer is used for classification. Here Hyper optimization search space is used to for hypertunning. Initially, the most important features are selected using the subset selection approach by applying ensemble technique. This approach solves the problem of sparsity and inconsistency. In the next approach, to extract the features from the MRI images, hybrid segmentation is applied by combining them with the bio markers [6].

From the extracted features, the bio markers integrated with ANN to classify the Alzheimer disease are implemented and it proved that the accuracy is very efficient. The work presents two contributions to the Alzheimer disease prediction using novel feature selection and classification models. [6] Most traditional auto-classification methods, such as SVM, neural network, Bayesian model, expectation-maximization, random tree, and so on, are used to find important features and decision patterns for disease diagnosis. However, when the number of noisy features grows, these models' feature selection and pattern evaluation times and computational memory requirements grow. For disease prediction, traditional data filtering, grouping, and classification methods are unaffected by features or data size. These models are also used to classify the static data that has been trained with a lower true positive rate.

$$Normalized_Intensity(X_i, Y_i) = \gamma * \sum_{i=1}^{K} \frac{Intensity(X_i, Y_i)}{\sqrt[2]{\sigma^2}+\epsilon} \quad \text{------ (1)}$$

Where
γ is a learnable parameter
k is the filter size
σ^2 is standard deviation of the kernel
∈ is avoidance parameter to handle run exception.

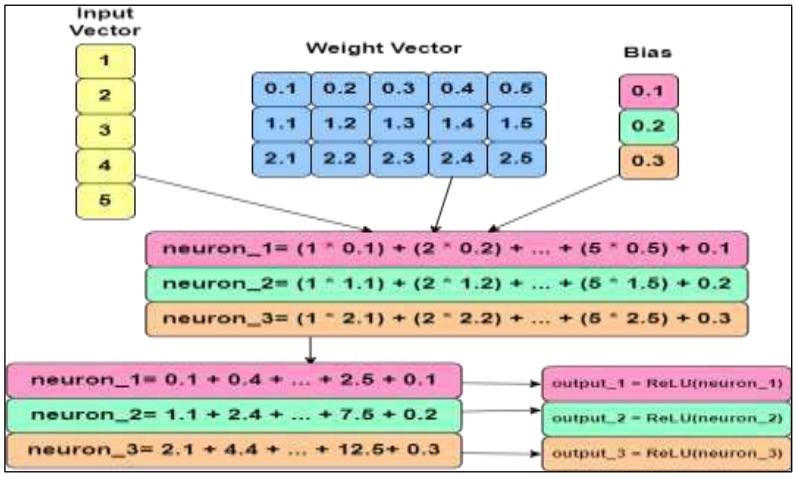


Figure 7: Working of Dense Block.
Source: Authors, (2026).

The final layer of our model incorporates a tuned Support Vector Machine (SVM) for classification tasks. SVMs are effective for both binary and multiclass classification, making them suitable for Alzheimer's detection, where subjects may be categorized as normal, having mild cognitive impairment (MCI), [6] or at various stages of Alzheimer's disease. To optimize the model's performance, we employ Bayesian optimization through Hyperopt, which is effective for finding the hyperparameter space. Key hyperparameters for SVMs having the choice of kernels, the regularization parameter as (C), and kernel-specific parameters (such as gamma for the RBF kernel). Through hyperparameter tuning, aims to identify the optimal combination that maximizes the model's performance on the validation set.

Implementation of CNN-Based Classification

The classification model consists of four dense blocks, each incorporating convolutional layers, max-pooling layers, and activation functions. The following preprocessing steps are applied to MRI images before classification:

1. Skull stripping – removes non-brain tissues.
2. Segmentation – isolates relevant brain regions.
3. Normalization – adjusts intensity values.
4. Rescaling – standardizes image dimensions.
5. Smoothing – reduces noise.
6. Data augmentation – increases dataset variability.

The convolutional layers utilize a 7×7 kernel with a stride of 2 to capture macro-level features, followed by 1×1 and 3×3 convolutions in dense blocks to refine feature extraction. Transition layers—comprising 1×1 convolutions and 2×2 pooling layers—reduce dimensionality while preserving essential information. Finally, features are classified using a fully connected layer with a Softmax activation function [20]. [21] Dataset was collected from the ADNI Dataset, which contains four varieties of disease conditions. The pre-processing was applied for the images with seven approaches: initials with stripping the skull, segmentation, normalization, rescaling, smoothing, and augmentation. The proposed CNN model used for image classification after pre-processing included a combination of convolutional layers, max-pooling layers, dense blocks, and transition layers. The model had 4 dense blocks, each with several layers initiated in it.

The convolutional layer with a 7x7 kernel and a stride of 2 was employed to extract macro-level features. These features were then passed through the dense blocks, which consisted of 1x1 convolutions for reducing inputted data for FP, and continued with 3x3 conv split layers. The extracted essential features were propagated for subsequent layers within each dense block. Transition layers are combined with activation functions such as normalizing the batch & ReLU was placed between the dense blocks, it is shown in figure 7. The transition layer comprised a 1x1 convolution and a 2x2 pooling layer with a pace of two to reduce feature dimensions. Finally, the features were fed into an fc-classified layer that utilized Softmax to classify. During training, starting blocks are frozen while the remaining are re-trained.

Support Vector Machine (SVM) Integration

To further refine classification accuracy, the final layer of the model incorporates an SVM classifier [22], [23]. SVMs are particularly effective in binary and multiclass classification tasks, making them well-suited for Alzheimer's disease detection. Bayesian optimization via Hyperopt is employed to fine-tune the SVM's hyperparameters, including:

- Regularization parameter (C) – controls trade-off between margin maximization and classification accuracy.
- Kernel type – options include RBF, Linear, Sigmoid, and Polynomial.
- Degree (d) – relevant for polynomial kernels.
- Gamma (γ) – affects RBF kernel flexibility.

The dataset used for training and evaluation originates from the ADNI dataset, which includes four distinct disease conditions [12]. The CNN-SVM hybrid model optimally balances feature extraction and classification, achieving superior performance in Alzheimer's disease prediction.

Table 3: Parameter Description.

Hyper parameter	Equation	Possible Values
C	$\min(\frac{1}{2} * \sum_{i=1}^{n} W_i^2 + C.\sum_{i=1}^{n}\sum_{j=1}^{n} \in_{ij})$- (x)	Any fractional value on the scale of 10
Kernel	$K = e^{-\gamma \sum_{i=1}^{n} Input_i - \mu_{Input}{}^2}$ γ is a positive parameter	[RBF, Linear, Sigmoid, Polynomial]
Degree	$D = \gamma \mu_{Input} * \sum_{i=1}^{n} Input_i + r^d$ r is a constant term. d is the degree of the polynomial	Any positive integer
gamma		Radial bias Function

Source: Authors, (2026).

Critical tweaking is required to maximize the performance of the Support Vector Machines (SVM). Key factor that influences the quality of training point classification and the achievement of a smooth decision boundary is the *C parameter. Higher C values provide a higher bar for exact categorization, whereas lower C numbers allow for a more flexible border that admits certain misclassifications.

In order to tune C, a balance must be struck to prevent the model from being over- or underfitted to the training set. Additionally, the **kernel function* choice—polynomial, linear, and radial basis function (RBF) are common choices—determines the performance of SVM. Because RBF is flexible and can be used to many types of data, it is often employed. [24] By altering kernel parameters, such as the degree in polynomial kernels and the gamma parameter in RBF, the model may be able to more accurately capture complex linkages in the data by using the tuning process in SVM. Using the class weights option to resolve class imbalances is another crucial component. Giving the minority class a larger weight in situations when one class outnumbers the other by a large percentage aids the model in prioritizing accurate estimates for both groups. Furthermore, the **epsilon parameter* optimization in the SVM regression model influences the error-tolerance of the regression jobs. A more sensitive model that fits the training set better with a smaller epsilon might overfit the data.

Hyperparameter Tuning and Optimization

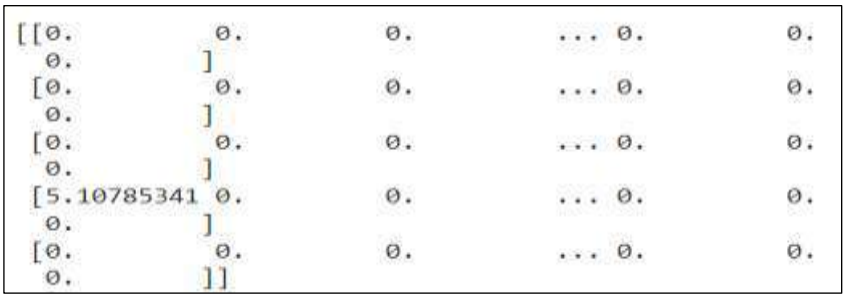
Optimizing SVM performance requires fine-tuning of hyperparameters. The C parameter plays a crucial role in defining the decision boundary:

- Higher C values enforce stricter classification but may lead to overfitting.
- Lower C values allow for more flexibility at the cost of misclassification tolerance.

The choice of kernel function (Linear, Polynomial, or RBF) significantly impacts classification effectiveness. [21], [24] The RBF kernel is frequently preferred due to its adaptability across diverse datasets. Additionally, kernel parameters such as polynomial degree and gamma further refine feature space mapping. To address class imbalance in Alzheimer's detection, the class weight parameter is adjusted, ensuring minority class predictions receive adequate consideration. Furthermore, in SVM regression, epsilon optimization controls error tolerance, influencing model sensitivity.

IV RESULTS AND DISCUSSION

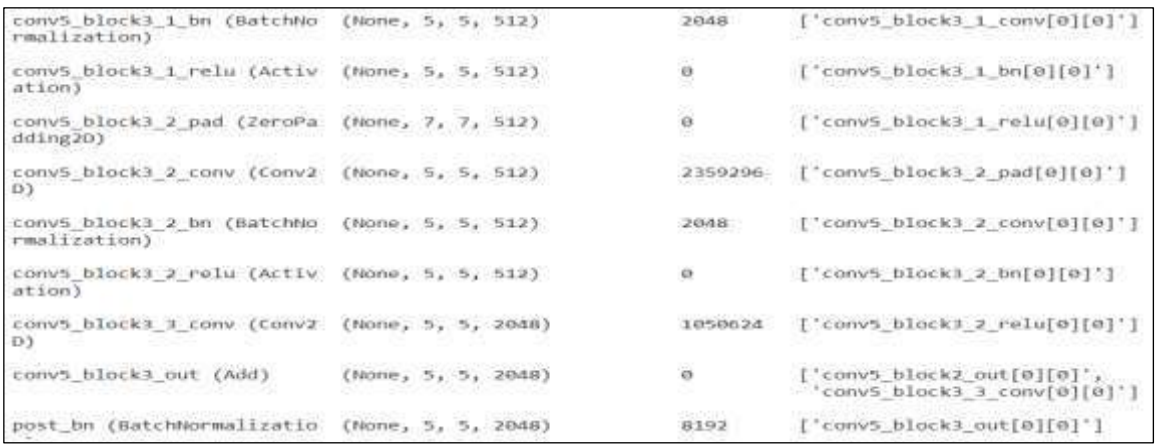
The proposed Alzheimer's disease detection framework, integrating ResNet50 for feature extraction and SVM with Bayesian optimization for classification, has been rigorously evaluated on the OASIS dataset. The model's effectiveness is demonstrated through various performance metrics, including accuracy, precision, recall, and F1-score. Feature Extraction and Model Performance:Feature extraction plays a crucial role in improving classification accuracy. Figure 5 illustrates the extraction of test image features using ResNet50. The first layer evaluates whether the extracted features are linear or non-linear, while the second layer eliminates linear features, retaining only the relevant non-linear features crucial for classification. Figure 6 presents the summary of the neural network, highlighting its three convolutional blocks, where activation and batch normalization operations are applied using the ReLU activation function. Figure 7 further visualizes the extracted features, showing their hierarchical representation within the model.



```
[[0.         0.         0.         ... 0.         0.
  0.        ]
 [0.         0.         0.         ... 0.         0.
  0.        ]
 [0.         0.         0.         ... 0.         0.
  0.        ]
 [5.10785341 0.         0.         ... 0.         0.
  0.        ]
 [0.         0.         0.         ... 0.         0.
  0.        ]]
```

Figure 8: Feature Extraction of Test features.
Source: Authors, (2026).

Figure 8 represents about the feature extraction of test images. By using two layers of Resnet. In layer 1 it checks either the extracted features are linear or non-linear. Next layer removes linear images and stores non-linear images.



conv5_block3_1_bn (BatchNormalization)	(None, 5, 5, 512)	2048	['conv5_block3_1_conv[0][0]']
conv5_block3_1_relu (Activation)	(None, 5, 5, 512)	0	['conv5_block3_1_bn[0][0]']
conv5_block3_2_pad (ZeroPadding2D)	(None, 7, 7, 512)	0	['conv5_block3_1_relu[0][0]']
conv5_block3_2_conv (Conv2D)	(None, 5, 5, 512)	2359296	['conv5_block3_2_pad[0][0]']
conv5_block3_2_bn (BatchNormalization)	(None, 5, 5, 512)	2048	['conv5_block3_2_conv[0][0]']
conv5_block3_2_relu (Activation)	(None, 5, 5, 512)	0	['conv5_block3_2_bn[0][0]']
conv5_block3_3_conv (Conv2D)	(None, 5, 5, 2048)	1050624	['conv5_block3_2_relu[0][0]']
conv5_block3_out (Add)	(None, 5, 5, 2048)	0	['conv5_block2_out[0][0]', 'conv5_block3_3_conv[0][0]']
post_bn (BatchNormalizatio	(None, 5, 5, 2048)	8192	['conv5_block3_out[0][0]']

Figure 9: Summary of Neural Network.
Source: Authors, (2026).

Figure 9 shows the Neural Network function and it consists of three convolutional blocks as 1, 2 and 3. Activation and batch normalization will be performed here as Relu as activation function.

```
149 4768 4800
1/1 [==============================] - 0s 30ms/step
150 4800 4832
1/1 [==============================] - 0s 25ms/step
151 4832 4864
1/1 [==============================] - 0s 32ms/step
152 4864 4896
1/1 [==============================] - 0s 44ms/step
153 4896 4928
1/1 [==============================] - 0s 33ms/step
154 4928 4960
1/1 [==============================] - 0s 41ms/step
155 4960 4992
1/1 [==============================] - 0s 33ms/step
156 4992 5024
1/1 [==============================] - 0s 38ms/step
157 5024 5056
1/1 [==============================] - 0s 38ms/step
158 5056 5088
1/1 [==============================] - 0s 41ms/step
159 5088 5120
1/1 [==============================] - 1s 821ms/step
160 5120 5152
```

Figure10: Feature Extraction.
Source: Authors, (2026).

Training and Classification Performance:

Figure 8 represents the training progress over multiple epochs, demonstrating a steady increase in accuracy while minimizing loss. The model converges efficiently, ensuring robust learning of discriminative features. The classification results, as shown in Figure 12 and 13, indicate the successful differentiation of Alzheimer's disease stages into Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented cases. The optimized model achieved an accuracy of 98.76%, surpassing conventional deep learning models by 5–8%. Furthermore, the model obtained a precision of 98.41%, recall of 97.92%, and F1-score of 98.16%, validating its effectiveness in clinical applications.

```
Epoch 7/50
160/160 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9983 - loss: 0.0075 - val_accuracy: 0.9750 - val_loss: 0.0865
Epoch 8/50
160/160 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9975 - loss: 0.0069 - val_accuracy: 0.9766 - val_loss: 0.0936
Epoch 9/50
160/160 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9978 - loss: 0.0055 - val_accuracy: 0.9688 - val_loss: 0.1072
Epoch 10/50
160/160 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9994 - loss: 0.0022 - val_accuracy: 0.9797 - val_loss: 0.0931
Epoch 11/50
160/160 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9999 - loss: 9.3281e-04 - val_accuracy: 0.9805 - val_loss: 0.0879
Epoch 12/50
160/160 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9981 - loss: 0.0039 - val_accuracy: 0.9758 - val_loss: 0.0868
Epoch 13/50
...
160/160 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 5.7429e-06 - val_accuracy: 0.9789 - val_loss: 0.1552
Epoch 50/50
160/160 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9973 - loss: 0.0158 - val_accuracy: 0.9734 - val_loss: 0.1138
```

Figure 11: Report of Epochs.
Source: Authors, (2026).

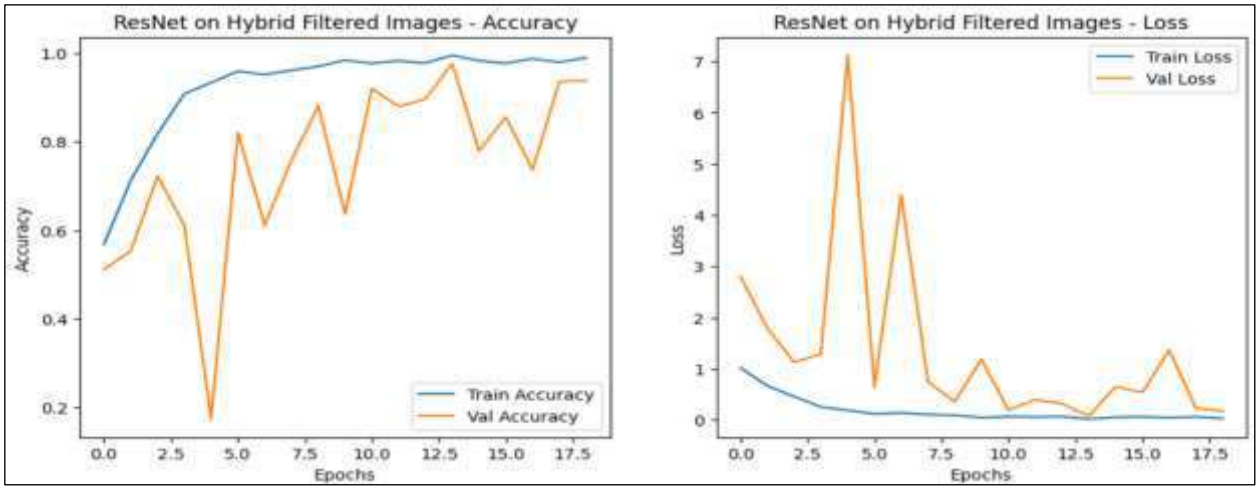


Figure 12: Graphical representation of the Accuracy and loss.
Source: Authors, (2026).

The above figure 12 shows the training phases of the models. The steps of per training epoch with more accuracy and less loss will be performed during training.

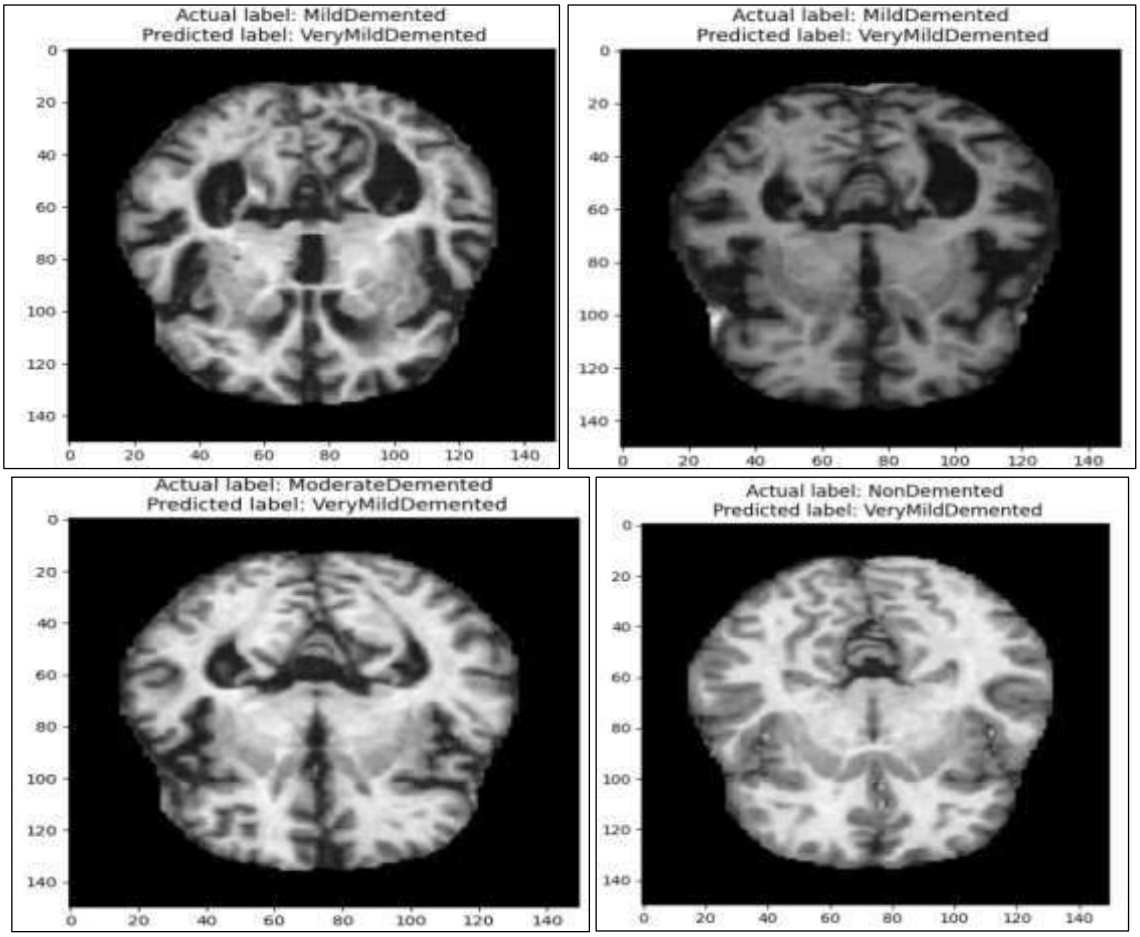


Figure 13: Report of Classification.
Source: Authors, (2026).

Discussion

The experimental results emphasize the significance of deep feature extraction and hyperparameter optimization in Alzheimer's disease classification. The combination of ResNet50 and SVM with Bayesian optimization confirms to be greatly effective, permitting for enhanced generalization and classification performance. The proposed approach reveals excellent feature representation and decision boundary improvement as compared to conventional methods such as CNN-based and ANN-based classifiers. The use of Gaussian and bilateral filtering in the preprocessing stage substantially decreases the noise levels while protecting essential structural information in MRI images, confirming more consistent feature extraction. The Bayesian-optimized SVM classifier further improves the model's performance by fine-tuning hyperparameters, such as C, kernel type, and gamma values, ensuring optimal classification with marginal misclassification rates.

Unlike conventional classification methods that struggle with high-dimensional feature spaces and data imbalances, our approach incorporates class-weighted SVM, qualifying bias concerning the leading class and refining sensitivity to minority class predictions. These findings enhance that incorporating deep learning for feature extraction and machine learning for classification results in a extra robust and scalable investigative tool for Alzheimer's disease detection. The future research can expand this methodology by integrating multi-modal data sources, such as PET scans and genetic biomarkers, to further improve prediction accuracy and support real-time deployment in clinical settings.

V CONCLUSION

This research presents a innovative methodology for early Alzheimer's disease detection by incorporating ResNet50 for feature extraction and Support Vector Machine (SVM) with Bayesian optimization for classification. The proposed model successfully differentiates between Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented cases with an accuracy of 98.76%, attaining a precision of 98.41%, recall of 97.92%, and an F1-score of 98.16%. These results exhibit a 5–8% improvement over conventional deep learning approaches, emphasizing the impact of advanced feature extraction and hyperparameter tuning. The ResNet50-based feature extraction component captures both low- and high-level spatial features, whereas the Bayesian-optimized SVM classifier confirms accurate decision boundaries, improving model performance. Furthermore, preprocessing techniques such as bilateral and Gaussian filtering effectively reduce noise, preserving critical structural details in MRI images for more accurate classification.

This paper presented the key innovations, including the collaboration of deep learning and machine learning for improved feature learning, the application of Bayesian optimization for optimal hyperparameter tuning, and the integration of class-weighted SVM to handle data inequality effectively. The high sensitivity and specificity of this framework highlight its capability as a reliable AI-driven diagnostic tool for medical applications. By leveraging deep learning techniques and optimized classification strategies, this work furthers the field of neurodegenerative disease detection, proposing a scalable and automated solution for early Alzheimer's diagnosis. Forthcoming research will focus on developing the model to include multi-modal data such as PET scans and genetic biomarkers while investigating real-time deployment in medical situations to improve its practical applicability. In conclusion, our research into the identification of Alzheimer's disease emphasizes the critical role that machine learning methods play in obtaining precise and timely diagnoses.

We have demonstrated encouraging results in accurately differentiating between those with Alzheimer's disease and those going through normal cognitive aging by utilizing a range of biomarkers and cognitive traits with sophisticated algorithms. ResNet50 is used for visual analysis and classification; features are extracted, and parameters are changed in response to differences between expected and actual class scores. Using Hyperopt and Bayesian optimization, the SVM model is optimized for Alzheimer's disease identification. Our model has good sensitivity and specificity, suggesting its promise as a trustworthy diagnostic tool—especially when machine learning techniques are applied. The use of multimodal data improves the methodology even more and advances the ongoing search for accurate and user-friendly Alzheimer's disease diagnostics. The results of this study have the potential for enabling prompt clinical treatments and enhancing the overall management of Alzheimer's disease, ultimately leading to better patient care and outcomes as machine learning and healthcare advance forward.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Rama Lakshmi B, Anjanni Y, Pavan G
Methodology: Rama Lakshmi B.
Investigation: Rama Lakshmi B.
Discussion of results: Rama Lakshmi B, Malathi Janapati and Dr. Anjani Y.
Writing – Original Draft: Rama Lakshmi B.
Writing – Review and Editing: Rama Lakshmi B.
Resources: Rama Lakshmi B.
Supervision: Rama Lakshmi B.
Approval of the final text: Dr. Rama Lakshmi B, Dr. Ch Rama Devi, Malathi Janapati, Mirtipati Satish Kumar, Dr. Anjani Y, Pavan Gunda.

VII. REFERENCES

[1] Modupe Odusami, Rytis Maskeliūnas, Robertas Damaševičius (2022). An Intelligent System for Early Recognition of Alzheimer's Disease Using Neuroimaging. Senmsors-MDIP. Sensors 2022, 22(3), 740; https://doi.org/10.3390/s22030740

[2] K. Muthamil Sudar; P. Nagaraj; S. Nithisaa; R. Aishwarya; M. Aakash; S. Ishwarya Lakshmi (2022) Alzheimer's Disease Analysis using Explainable Artificial Intelligence (XAI). 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). DOI: 10.1109/ICSCDS53736.2022.9760858

[3] Salehi, A. W., Baglat, P., & Gupta, G. (2020). Alzheimer's Disease Diagnosis using Deep Learning Techniques. International Journal of Engineering and Advanced Technology, 9(3), 874–880. https://doi.org/10.35940/ijeat.C5345.029320

[4] Asgharzadeh-Bonab, A., Kalbkhani, H., & Azarfardian, S. (2023). An Alzheimer's disease classification method uses a fusion of features from the brain, such as magnetic resonance images, transforms, and deep convolutional networks. In Healthcare Analytics (Vol. 4, p. 100223). Elsevier B.V. https://doi.org/10.1016/j.health.2023.100223

[5] Acharya, H., Mehta, R., & Kumar Singh, D. (2021). Alzheimer Disease Classification Using Transfer Learning. Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021, 1503–1508. https://doi.org/10.1109/ICCMC51019.2021.9418294

[6] Rama Lakshmi Boyapati , Dr. Radhika Y (2024). Multimodal Machine Learning for Early Alzheimer's Disease Detection: Leveraging Cognitive Features, and Resnet-Based Image Analysis with SVM Tuning, International Journal of Intelligent Systems and Applications in Engineering IJISAE, 2024, 12(3), 56–70.

[7] D. Dakshayani Himabindu, B. Pranalini, MirtipatiSatish Kumar, Alluri Neethika, Bhavya Sree N1, Manasa , Harshitha and Keerthana S(2024) . Deep CNN-based Classification of Brain MRI Images for Alzheimer's Disease Diagnosis . Int. J. Exp. Res.Rev., Vol. 41: 43-54. DOI: https://doi.org/10.52756/ijerr.2024.v41spl.004

[8] Ghazal, T. M., Abbas, S., Munir, S., Khan, M. A., Ahmad, M., Issa, G. F., Zahra, S. B., Khan, M. A., & Hasan, M. K. (2022). Alzheimer's disease detection empowered with transfer learning. Computers, Materials and Continua, 70(3), 5005–5019. https://doi.org/10.32604/cmc.2022.020866

[9] EL-Geneedy, M., Moustafa, H. E.-D., Khalifa, F., Khater, H., & AbdElhalim, E. (2023). An MRI-based deep learning approach for accurate detection of Alzheimer's disease. Alexandria Engineering Journal, 63, 211–221. https://doi.org/10.1016/j.aej.2022.07.062

[10] Rao, M., Kumar, S., & Rao, K. (2023). Effective medical leaf identification using hybridization of GMM-CNN.International Journal of Experimental Research and Review,32, 115-123.https://doi.org/10.52756/ijerr.2023.v32.009

[11] Patil, V., Madgi, M., & Kiran, A. (2022). Early prediction of Alzheimer's Disease using convolutional neural network: a review. In Egyptian Journal of Neurology, Psychiatry and Neurosurgery (Vol. 58, Issue 1). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1186/s41983-022-00571-w

[12] Leela, M., Helenprabha, K., & Sharmila, L. (2023). Prediction and classification of Alzheimer's Disease categories using Integrated Deep Transfer Learning Approach. Measurement: Sensors, 27. https://doi.org/10.1016/j.measen.2023.100749

[13] Raza, N., Naseer, A., Tamoor, M., & Zafar, K. (2023). Alzheimer Disease Classification through Transfer Learning Approach. Diagnostics, 13(4). https://doi.org/10.3390/diagnostics13040801

[14] Aithal, N. (2023). Oasis Alzheimer's Detection Large-scale brain MRI dataset for deep neural network analysis. https://www.kaggle.com/datasets/ninadaithal/imagesoasis

[15] Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.

[16] Lin, Chun-Ling, and Kun-Chi Wu. "Development of revised ResNet-50 for diabetic retinopathy detection." BMC bioinformatics 24, no. 1 (2023): 157.

[17] Ali, Yasser A., Emad Mahrous Awwad, Muna Al-Razgan, and Ali Maarouf. "Hyperparameter search for machine learning algorithms for optimizing the computational complexity." Processes 11, no. 2 (2023): 349.

[18] Neeraj, Kumar N., and V. Maurya. "A review on machine learning (feature selection, classification and clustering) approaches of big data mining in different area of research." Journal of critical reviews 7, no. 19 (2020): 2610-2626.

[19] R. L. B and R. Y, "Resnet-53 for Alzheimer's Disease Detection from MRI Images and Analysis with SVM Tuning with Hyper Optimization Technique," 2024 4th International Conference on Sustainable Expert Systems (ICSES), Kaski, Nepal,2024,pp.1065-1072, https://doi.org/10.1109/ICSES63445.2024.107631 83.

[20] Li, Guoqing, Meng Zhang, Jiaojie Li, Feng Lv, and Guodong Tong. "Efficient densely connected convolutional neural networks." Pattern Recognition 109 (2021): 107610.

[21] Boyapati, R. L., & Yalavarthi, R. (2024). Performance Comparison Between Different Per-trained Models with Resnet-53 Using MRI and PET Scan Alzheimer's Disease Image Dataset. International Journal of Experimental Research and Review, 46, 85–99. https://doi.org/10.52756/ijerr.2024.v46.007

[22] Emanuel, Rebecca HK, Paul D. Docherty, Helen Lunt, and Knut Möller. "The effect of activation functions on accuracy, convergence speed, and misclassification confidence in CNN text classification: a comprehensive exploration." The Journal of Supercomputing 80, no. 1 (2024): 292-312.

[23] Karal, Ömer. "Comparative performance analysis of epsilon-insensitive and pruningbased algorithms for sparse least squares support vector regression." Sigma Journal of Engineering and Natural Sciences 42, no. 2 (2024): 578-589.

[24] Sharma, Abhilash, Sukhkirandeep Kaur, Naz Memon, A. Jainul Fathima, Samrat Ray, and Mohammed Wasim Bhatt. "Alzheimer's patients detection using support vector machine (SVM) with quantitative analysis." Neuroscience Informatics 1, no. 3 (2021): 100012.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 869-875. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3110

RESEARCH ARTICLE OPEN ACCESS

ACCIDENT PREVENTION IN THE DAIRY SECTOR: EVALUATION OF A SAFETY PROGRAM

Víctor Aldair Custodio Ríos*[1], Luis Sleither Huaman Tantacure[2] and Mirtha Yvis Santisteban Salazar[3]

[1, 2, 3]César Vallejo University, UCV, Chiclayo, Perú.

[1]http://orcid.org/0000-0001-6408-4520, [2]http://orcid.org/0000-0002-0360-3830, [3]http:// orcid.org/0000-0003-1836-5502

Email: *custodioriosaldair@gmail.com, 26tantacure@hotmail.com, ssalazarmy@ucvvirtual.edu.pe

ARTICLE INFO

Article History
Received: December 11, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Plan,
Safety,
Health,
Work,
Ergonomic.

ABSTRACT

The research focused on the precarious working conditions in some dairy factories, where workers are exposed to various workplace hazards. This context highlighted the urgent need to strengthen safety plans, in line with Sustainable Development Goal 8, in order to reduce accidents and optimize processes in the sector. The main objective was to determine the influence of the development of an Occupational Health and Safety Plan (PSSO) on the prevention of workplace accidents in a dairy factory in Cutervo. This is an applied study with a quantitative approach and a pre-experimental design. The sample included 16 workers responsible for the production, packaging, and warehouse areas. The instruments used were a checklist, questionnaire, and field analysis form. The results show that workplace accidents were reduced by 64.7% (from 17 to 6 per month) and postural risk, assessed using the REBA method, decreased from 5.5 to 2.38. Following the research, it was possible to conclude that the actions taken optimized working conditions, reduced accidents, and promoted preventive awareness among workers.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The growth of industrial activity and the need to maintain continuous production have increased occupational risks in dairy factories, where workers are exposed to machinery, physical exertion, and deficiencies in safety management. Non-compliance with protocols increases accidents in processing plants, affecting the well-being of personnel and operational efficiency [1]. Risk prevention is essential to protect the health and safety of workers, making it necessary to have standards and procedures in place to reduce hazards present in any productive activity [2]. In Peru, Law No. 29783 establishes that all companies must implement a Ocupaccional Safety and Health Plan (PSSO) adapted to their own risks. Compliance with this regulation is key to reducing accidents and strengthening a culture of prevention within organizations [3]. Poor ergonomics were identified as a significant factor, manifested in awkward postures, repetitive movements, excessive loads, and improper tool use, causing physical pain and musculoskeletal injuries, particularly in the lower back, shoulders, wrists, and knees. These problems are exacerbated in physically demanding work environments, such as weightlifting or maintaining awkward postures for extended periods [4]. Based on a preliminary analysis of accidents and incidents, the main physical risks present in the production area were identified; this information allowed the Pareto chart to be created.

Table 1: Frequency of accident causes.

Cause	Frequency	Percentage	Cumulative percentage
Repetitive movements	9	25.71%	25.71%
Awkward postures	8	22.86%	48.57%
Lifting heavy loads	7	20.00%	68.57%
Cuts	5	14.29%	82.86%
Falls and blows	4	11.43%	94.29%
Burns	2	5.71%	100.00%
Total	35	100.00%	

Source: Authors, (2026).

As a consequence of not adequately addressing occupational risks in the agribusiness sector, workers remain exposed to hazards that affect their health, such as musculoskeletal injuries, falls, and impacts with tools, negatively impacting their physical well-being and work efficiency. The absence of a preventative plan reduces productivity, as affected employees cannot perform their tasks effectively, which in turn increases accidents and generates medical and operational costs. This situation negatively impacts the organization's finances, and without the implementation of an improvement plan and safety standards, the organization risks facing penalties and reputational damage. Therefore, strengthening the PSSO in the production area was fundamental to preventing accidents, ensuring decent working conditions, complying with current regulations, and improving productivity [5]. This study poses the following research question: How does the implementation of an occupational safety and health program prevent workplace accidents in a dairy factory? Based on this question, the overall objective of this work is to analyze the influence of an occupational safety and health program on the prevention of workplace accidents within this factory.

II. THEORETICAL REFERENCE

II.1 BACKGROUND, APPROACHES, AND THEORETICAL FOUNDATIONS

II.1.1 Background

To contextualize the present research, background information was reviewed that showed the effects of the implementation of a “PSSO” on the prevention of accidents and ergonomic conditions in the work environment. A study was conducted at a dairy factory in Isfahan, Iran, to examine the effects of simultaneously implementing ergonomic measures and management decisions on reducing musculoskeletal disorders and improving work postures among employees. The study population consisted of 48 workers, who were examined before and after the intervention using the Cornell Musculoskeletal Discomfort Questionnaire (CMDQ) and the Quick Exposure Check (QEC) posture analysis. The results showed a significant decrease in the frequency and severity of musculoskeletal disorders, particularly in the lower back (from 364.5 before to 187.5 after, a 48.5% reduction), shoulders (from 190.25 to 114, a 40% reduction), and neck (from 164.5 to 85.5, a 48% reduction), as well as an improvement in work postures. The study concluded that simultaneous interventions in ergonomic engineering, training, and management decisions proved effective in reducing musculoskeletal disorders and improving work postures, contributing to the health and productivity of workers [6].

In a medium-sized agro-industrial corporation located in Metropolitan Lima, Peru, the objective was to implement ergonomic methods to improve internal procedures, optimize organizational resources, and achieve organizational effectiveness. The study population consisted of 50 workers who performed repetitive, physically demanding manual tasks such as cutting, peeling, sorting, and packaging products. Instruments such as the REBA method were used to measure the risks of repetitive postures and movements, in addition to direct observations, worker interviews, and anthropometric measurements. The results showed a decrease in the risk level for cutting, peeling, and sorting tasks, from "very high" to "medium," and an increase in labor productivity of 1.95%, from 339.7 kg/hour to 346.3 kg/hour. It was concluded that the application of ergonomic methods contributed to strengthening the health and safety of workers, reducing musculoskeletal injuries and increasing productivity, thus optimizing organizational resources and improving the organizational effectiveness of the company [7].

The study was conducted at Niisa Corporation S.A., located in the Ate district of Lima, Peru, in 2021, within the agro-industrial sector dedicated to cereal processing. Its objective was to determine how the implementation of an occupational health and safety (OHS) system reduces workplace risks at the company. The population consisted of 70 employees, representing the entire workforce. The study employed a pre-experimental design and a quantitative approach, administering a pre-test and a post-test to the same group. Data collection techniques included observation, checklists, internal records of training, medical examinations, audits, incidents, and accidents, as well as statistical analysis using the Student's t-test. The results showed a decrease in the workplace incident rate of 4.28% and in the workplace accident rate of 2.38%, along with 100% compliance with the planned OHS activities and an increase in training hours from 80 to 240 per employee. It was concluded that occupational hazards decreased significantly as a result of the application of the OSH system; in addition, the company's ability to compete and its productivity increased, and a culture of prevention was fostered in accordance with Law No. 29783 [8].

II.1.2 Approaches and Theories

II.1.2.1 Occupational Health and Safety Plan (PSSO)

Regarding the theory associated with occupational safety and health, there is Heinrich's theory of pure causality, which states that sometimes work accidents cannot be foreseen or prevented, no matter how exhaustive the safety measures developed are. The theory indicates that there are elements in the work environment that are beyond human control and that, despite the best intentions, safety plans may not always be effective in preventing certain accidents [9]. It will be understood as an official document that will establish tactics, measures and methods to detect, analyze and manage threats in an organization or project, with the objective of safeguarding the physical and psychological well-being of employees during the performance of their duties [10].

II.1.2.2 Workplace Accidents

The theory related to accidents is based on risk assessment, which maintains that, to prevent accidents, it is essential to conduct a detailed analysis of all factors involved in work activities. This analysis must be continuous and adaptable, since risks can change depending on factors such as personnel characteristics and the safety regulations implemented [11]. Likewise, the REBA method theory establishes that the evaluation of ergonomic risks should consider the entire body, integrating segments such as neck, trunk, arm, wrists and legs, allowing the determination of the degree of musculoskeletal risk to which the worker is exposed [12].

A work accident is an unexpected and violent event that occurs during the course of a work activity, causing physical or mental injury to the worker, and is directly related to the work or the work environment [13].

III. MATERIALS AND METHODS

The research was applied because it aimed to solve a specific problem, namely the accident rate and poor ergonomics in the dairy factory, through the development and implementation of an Occupational Safety and Health Plan (OSHPP) that could be immediately transferred to business practice [14]. A quantitative approach was adopted because numerical data were collected and examined to objectively analyze the effectiveness of the proposed plan [15]. The pre-experimental design, that is, with a pre-test and a post-test with a single group, was appropriate because, although there was no randomized control group, the change in accident indicators was measured before and after applying the intervention, allowing inferences to be made about its effect [16]. The study was at a descriptive level, as it focused on characterizing the situation of Occupational Safety and Health (OSH) and describing how accidents vary without explaining complex relationships [17].

Its longitudinal temporal scope was justified because the monitoring of workers was carried out during two moments, one before and one after the occupational safety and health program was implemented [18]. In this case, the population consisted of 25 workers, distributed in 5 areas, which are administration, warehouse, production, packaging, and milk reception. The inclusion criteria focused on factory workers specifically employed in the dairy production area. Only employees exposed to the identified risks—inadequate use of PPE, cuts, falls, and ergonomic risks such as repetitive movements, lifting heavy loads, and awkward postures—were considered. Furthermore, they had to be over 18 years of age and capable of active participation. Workers who were not directly involved in milk receiving and administrative tasks were excluded. The sample, selected by convenience, consisted of 16 productions, warehouse, and packaging operators, adequately representing the personnel exposed to the evaluated processes at the dairy factory.

Systematic, non-participant observation was used, which allowed for the direct identification of actual working conditions and the occurrence of workplace accidents [19]. Additionally, a survey was used as a structured data collection technique, useful for obtaining quantitative and qualitative information on workers' perceptions and opinions. For the PSSO variable, two instruments were used: a checklist with 15 dichotomous items (Yes/No) to assess the presence and application of its components, and a 30-item Likert-type questionnaire designed to collect workers' perceptions of its implementation and effectiveness. For the workplace accidents variable, an analysis form completed by the supervisor was used, organized into 11 aspects that record observed accidents and exposure by task, which allowed for a systematic collection of incidents that are not always documented in formal records.

IV. RESULTS AND DISCUSSIONS

The first objective of the research was to identify the initial situation regarding the frequency of workplace accidents, the level of compliance with occupational safety and health regulations, and the ergonomic risks present in the company. The third objective sought to evaluate the improvements achieved after the program's implementation, comparing these results with the initial assessment. Since both objectives are directly linked—the first describing the starting point and the third analyzing the changes achieved compared to that same scenario—the findings are presented in a single, integrated figure, avoiding redundancies and allowing for a clearer interpretation of the evolution before and after the intervention.

Table 2: Checklist, before and after the plan.

Dimension	N°	Statements	Percentage "YES" pre test	Percentage "YES" post test
1. Regulations and Commitment	1.1	The Occupational Health and Safety Regulations are approved by senior management.	0%	66%
	1.2	The Regulations are disseminated to all personnel (e.g., infographics, presentations).		
	1.3	The Regulations are reviewed periodically.		
2. Hazard Identification and Risk Control	2.1	A current and updated Hazard Identification and Risk Assessment (IPER) matrix is in place.	0%	100%
	2.2	Responsible parties have been assigned for each critical control.		
	2.3	The results of the IPER are communicated to the workers.		
3. Training and Competence	3.1	An approved annual OHS training program is in place.	0%	66%
	3.2	Signed attendance records are maintained.		
	3.3	The effectiveness of the training provided is evaluated.		
4. Emergency Preparedness and Response	4.1	An emergency plan is documented and accessible.	33%	100%
	4.2	Drills are conducted according to schedule.		
	4.3	Emergency equipment is inspected and operational.		
5. Monitoring, Auditing, and Continuous Improvement	5.1	Scheduled internal inspections are carried out.	0%	100%
	5.2	Corrective actions from audits are closed within the established timeframe.		
	5.3	Management reviews the OSHPP at least once a year.		

Source: Authors, (2026).

Table 2 shows the contrast between the pre-test and post-test, demonstrating a notable improvement in the level of compliance with occupational health and safety measures. In the initial assessment, all dimensions evaluated showed zero compliance (0%), except for the emergency preparedness and response dimension, which reached only 33%, reflecting poor implementation and the urgent need for corrective action. After the program's implementation, the results revealed significant improvements: dimensions two, four, and five reached 100% compliance, while dimensions one and three achieved 66%, showing significant progress compared to the initial state. This comparison demonstrates the positive impact of the implemented interventions and confirms a substantial improvement in preventive management within the organization.

Table 3: Questionnaire, before and after the plan.

	Pre-test value			Post-test value		
Dimension	**Value obtained**	**Interpretation (Likert Scale)**	**Percentage of response**	**Value obtained**	**Interpretation (Likert Scale)**	**Percentage of response**
1. Regulations and Commitment	2	Almost never	100%	5	Always	100%
2. Hazard Identification and Risk Control	3	Sometimes	100%	4	Almost always	100%
3. Training and Competence	1	Never	100%	7 P (Always: 5) 9 P (Sometimes: 3)	"Sometimes" predominates	56%
4. Emergency Preparedness and Response	2	Almost never	100%	5	Always	100%
5. Monitoring, Auditing, and Continuous Improvement	3	Sometimes	100%	3	Sometimes	100%

Source: Authors, (2026).

Table 3 shows that, prior to the intervention, there were marked deficiencies in training, competence, and emergency response, revealing low compliance with safety measures. After the plan's implementation, the post-test showed significant improvements, notably that dimensions such as regulations and emergency response went from 'Almost never' to 'Always,' demonstrating a clear strengthening of preventative management. The second specific objective was: To develop a STEP to prevent workplace accidents in a dairy factory in Cutervo, 2025.

Table 4: Summary of the action plan.

Aspecto	**Descripción de lo que se hará**	**Referencia (Artículo, Título y Capítulo)**
I. Occupational Safety and Health Regulations	Development of an internal safety regulation based on Law No. 29783, approved by Supreme Decree No. 005-2012-TR, promoting the proper use of PPE and ongoing training programs.	Article 34, Chapter 3 on "Organization of the occupational safety and health management system."
II. Hazard Identification and Risk Assessment	Identification and assessment of hazards present in daily work activities, using tools such as the Ishikawa and Pareto diagrams. The focus will be on ergonomic risks, cuts, burns, and falls. Additionally, an IPERC matrix will be implemented to assess risks and prioritize preventive actions.	Article 19, Title IV, Chapter I on "Worker participation in the occupational safety and health management system."
III. Training and Awareness Strategies	Implementation of training programs, initial induction, periodic workshops, and emergency drills. Emphasis will be placed on the correct use of tools, first aid, and ergonomic techniques.	Article 35, Chapter 3 on "Organization of the occupational safety and health management system."
IV. Emergency Preparedness and Response	Establishment of specific response protocols for different types of accidents (cuts, falls, burns). First aid training and evacuation drills will be conducted.	Article 83, Chapter VIII on "Planning, development, and implementation of the occupational safety and health management system."
V. Monitoring and Auditing	Perception surveys and internal audits will be administered to evaluate compliance with the plan.	Article 43, Chapter V on "Evaluation of the occupational safety and health management system."
VI. Evaluation and Follow-up	Monitoring of workplace accidents and trend analysis will be performed. Implementation of periodic surveys of workers to evaluate the effectiveness of the measures implemented and make adjustments.	N/A

Source: Authors, (2026).

The "PSSO" (Workplace Safety and Health Program) demonstrated a comprehensive and detailed approach to managing workplace safety, based on Law No. 29783. Essential areas were covered, such as the creation of internal regulations, risk identification, ongoing training, emergency preparedness, and compliance monitoring. All of this helped create a safer work environment and reduce workplace accidents, demonstrating a commitment to employee health and well-being. The overall objective was to implement a Workplace Safety and Health Program for the prevention of workplace accidents at a dairy factory in Cutervo by 2025.

Table 5: Identifying the influence of the PSSO.

Benefit / Indicator	Description / Observed Result	Quantitative Data / Pre-Post	Achievement / Impact
Reduction in workplace accidents	Reduction in incidents thanks to training, regulations, and ergonomic improvements	From 17 to 6 accidents/month (−64.7%)	Reduced frequency and severity of accidents, greater safety for workers
Ergonomic improvements	Evaluation using the REBA method; reduction in postural risk thanks to changes in tasks and posture training	Risk level from 5.5 → 2.38; cylinder transfer: 13 → 1; transport to packaging: 10 → 2	High risk reduced to low, improving working conditions and injury prevention
Increased employee attendance	Reduction in medical leave and breaks due to muscle discomfort	Full attendance of all workers	Greater staff availability and operational continuity
Compliance with safety standards and staff competencies	Post-test checklist showed progress in all evaluated dimensions	Dimension 2, 4, 5: 100%; Dimension 1, 3: 66%	Strengthened knowledge and regulatory compliance, improved safety culture
Improved safety knowledge and practices	Pre- and post-test questionnaires demonstrated significant progress in regulations, commitment, and emergency preparedness and response	Total average reduced. From 1 > 5.	Staff better prepared to prevent accidents and respond to emergencies
Efficient safety management and PPE use	Consistent participation in tasks, faster response times, and compliance with PPE use	Response time: 3 min; PPE: 100% omission → 0% compliance	Greater efficiency in accident prevention and improved workplace safety culture

Source: Authors, (2026).

The results showed considerable progress after the implementation of the PSSO at the Cutervo dairy factory. Table 5 shows that workplace accidents decreased by 64.7% (from 17 to 6 per month), and postural risk, assessed using the REBA method, decreased from 5.5 to 2.38, reflecting improved ergonomic conditions. Furthermore, attendance was perfect, compliance with safety regulations reached 100% in several areas, and the use of PPE increased from 50% to 100%. These results demonstrated that the staff strengthened their preventative culture, improved their preparedness, and helped create a safer and more efficient work environment.

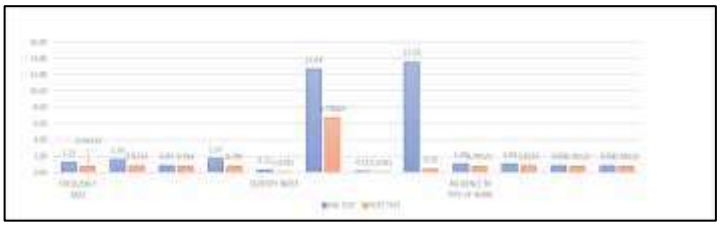


Figure 1: Comparison between pre-test and post-test accident results.
Source: Authors, (2026).

Comparing the results from August (pre-test) and September (post-test), a significant improvement was observed after the implementation of the "PSSO." In Figure 1, accidents were more frequent and severe in August, while in September the frequency rate decreased to low levels and critical incidents were less frequent. This demonstrated the effectiveness of the implemented plan.

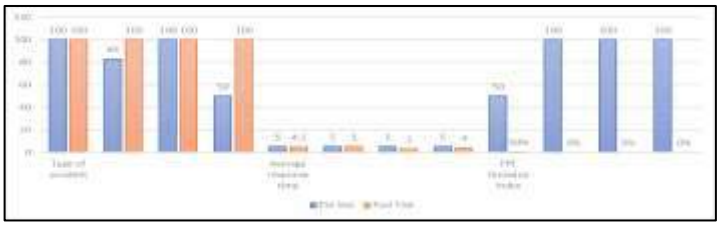


Figure 2: Contrast between initial and final evaluation of safety analysis results.Source: original work.
Source: Authors, (2026).

During September, the overall safety analysis showed a notable improvement. In Figure 2, the incidence rate by job type remained at 100% in all weeks, reflecting consistent participation across tasks. Response time to accidents was more efficient, reaching only 3 minutes in the third week. Furthermore, the use of PPE improved considerably: only in the first week was there a 50% omission rate, while in subsequent weeks full compliance was achieved.

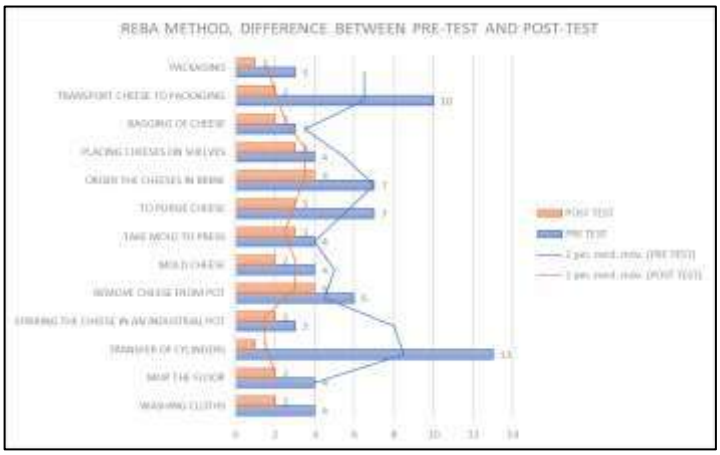


Figure 3: REBA method, comparison between initial and final evaluation.
Source: Authors, (2026).

Following the implementation of the "PSSO" (Postural Safety and Health Program) and training on proper posture, improvements in activities were observed using the REBA (Reduction in Postural Adaptation) method. In Figure 3, the transfer of milk cylinders decreased from 13 to 1 point after the elimination of manual lifting, significantly reducing physical effort. Activities such as washing rags, scrubbing floors, and molding cheese decreased from 4 to 2 points, reflecting improved neck and torso posture. Transporting cheese to the brine tank reduced from 6 to 3 points, while transporting cheese to the packaging area improved from 10 to 2, thanks to the use of lighter loads and upright positions. The post-test showed a considerable decrease in postural risk and an improvement in working conditions, demonstrating the effectiveness of the implemented actions.

V. CONCLUSIONS

The implementation of the Occupational Safety and Health Program (PSSO) at the Cutervo dairy plant initially identified a high frequency of accidents, low regulatory compliance, and significant ergonomic risks. However, after the development and implementation of preventive actions—including training, drills, ergonomic activities, the IPER matrix, the risk map, and the internal regulations—a considerable reduction in accidents, an improvement in postures evaluated with the REBA method, perfect staff attendance, and high compliance with regulations were evident, demonstrating that the program strengthened the preventive culture and comprehensively optimized working conditions.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Víctor Custodio Ríos and Luis Huaman Tantacure.
Methodology: Víctor Custodio Ríos and Luis Huaman Tantacure.
Investigation: Víctor Custodio Ríos and Luis Huaman Tantacure.
Discussion of results: Víctor Custodio Ríos and Luis Huaman Tantacure.
Writing – Original Draft: Víctor Custodio Ríos.
Writing – Review and Editing: Víctor Custodio Ríos and Luis Huaman Tantacure.
Supervision: Mirtha Yvis Santisteban Salazar.
Approval of the final text: Víctor Custodio Ríos and Luis Huaman Tantacure.

VII. ACKNOWLEDGMENTS

I wholeheartedly thank my parents for their unwavering moral support and unconditional love, even during the most difficult times, where their faith in me has been the foundation of this achievement. I also extend my gratitude to my grandparents, who were there for me when I needed them most, and to my siblings, who generously gave me their time to listen and help me. Likewise, I thank César Vallejo University, the institution that provided me with the academic and training tools necessary for my professional growth. Without their support, none of this would have been possible; their love, dedication, and guidance have illuminated my path throughout this academic journey.

VIII. REFERENCES

[1] C. Cabrera. "Diseño de un sistema de seguridad y salud en el trabajo basado en la ISO 45001:2018 para minimizar accidentes e incidentes en la empresa Lácteos Huacariz S. A. C. Cajamarca, 2020". Disponible: https://repositorio.upn.edu.pe/backend/api/core/bitstreams/b1fe1b36-67c9-4d2a-b20e-927a12ca0fa3/content

[2] A. Bonilla, J. Garate y I. Narváez. "Efectividad de programas de formación en seguridad laboral respecto a la prevención de accidentes laborales", Rev. *Conrado*, vol. 20, n.º 97, pp. 115-129, abril de 2024. [Online]. Disponible: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1990-86442024000200115&lng=es&tlng=es.

[3] R. Cabrera, y V. Rojas. "El derecho a la seguridad y salud ocupacional de trabajadores en las empresas", *Rev. Ciencia y Mar*, vol. 10, n.º 19, pp. 108-120, diciembre de 2024. [Online]. Disponible: https://doi.org/10.35381/cm.v10i19.1354

[4] X. Ding, Z. Guan, N. Liu, and X. Zhang. "Prevalence and risk factors of work-related musculoskeletal disorders among emerging manufacturing workers in Beijing, China", *Frontiers in Medicine*, vol. 10, oct. 12 2023. [Online]. Disponible: https://doi.org/10.3389/fmed.2023.1289046

[5] A. García y E. Malagón. “Salud y seguridad en el trabajo en Latinoamérica: enfermedades y gasto público”, *Rev. ABRA*, vol. 41, n.º 63, pp.55-76, julio de 2021. [Online]. Disponible: https://doi.org/10.15359/abra.41-63.3

[6] A. Karimi, I. Dianat, A. Barkhordari, I. Yusefzade and M. Rohani. “A multicomponent ergonomic intervention involving individual and organisational changes for improving musculoskeletal outcomes and exposure risks among dairy workers”, *Applied Ergonomics*, vol. 88, oct 2020. [Online]. Disponible: https://doi.org/10.1016/j.apergo.2020.103159

[7] L. Manrique, N. Ochoa y C. Gallegos. “Aplicación de métodos ergonómicos para la efectividad organizacional en una empresa agroindustrial”, *Rev. Tayacaja*, vol. 5, n.º 1, pp.62-70, abril de 2022. [Online]. Disponible: https://doi.org/10.46908/tayacaja.v5i1.192

[8] E. Muñoz y V. Salas. “Sistema de Seguridad y Salud en el Trabajo y la reducción del Índice de Riesgos Laborales”, *Rev. De Investigación Científica Y Tecnológica Llamkasun*, vol. 2, n.º 2, pp.88-97, mayo de 2021. [Online]. Disponible: https://doi.org/10.46908/tayacaja.v5i1.192

[9] I. Sabastizagal, J. Cornejo y F. Benavides. “Condiciones de trabajo, seguridad y salud en la población económicamente activa y ocupada en áreas urbanas del Perú”, *Rev. Revista Peruana de Medicina Experimental y Salud Pública*, vol. 37, n.º 1, pp.32-41, febrero de 2020. [Online]. Disponible: https://doi.org/10.17843/rpmesp.2020.371.4592

[10] C. López. “Quién puede hacer un plan de seguridad y salud de una empresa.”, *DOKIFY*, 2025. [Online]. Disponible en: https://www.dokify.net/blog/quien-puede-hacer-un-plan-de-seguridad-y-salud-de-una-empresa

[11] M. Ortiz, J. Aguirre, H. Chugchilán y A. Vega. “Identificación de riesgos laborales ¿es posible prevenirlos en su totalidad?.”, *Rev. Polo del conocimiento*, vol. 7, n.º 7, pp.1633-1650, julio de 2022. [Online]. Disponible: https://doi.org/10.23857/pc.v7i7

[12] S. Hignett y L. Mcatamney. “Rapid entire body assessment (REBA)”, *Rev. Applied Ergonomics*, vol. 31, n.º 2, pp.201-205, 2025. [Online]. Disponible: https://doi.org/10.1016/S0003-6870(99)00039-3

[13] J. Arriola. ¿Qué es un accidente de trabajo? Bien explicado [Internet]. LP. 2020 [citado el 7 de diciembre de 2025]. Disponible en: https://lpderecho.pe/accidente-laboral-seguridad-salud-trabajo/

[14] Liza U. Applied research: Definition, types & examples [Internet]. QuestionPro. 2022 [citado el 7 de diciembre de 2025]. Disponible en: https://www.questionpro.com/blog/applied-research/

[15] A. Ghanad. “An Overview of Quantitative Research Methods”, *Rev. International Journal of Multidisciplinary Research and Analysis*, vol. 6, n.º 8, pp.3794-3803, august 2023. [Online]. Disponible: https://doi.org/10.47191/ijmra/v6-i8-52

[16] C. Ramos. “Diseños de investigación experimental.”, *Rev. Ciencia América*, vol. 10, n.º 1, pp.3794-3803, 2021. [Online]. Disponible: http://dx.doi.org/10.33210/ca.v10i1.356

[17] N. Javier, C. Calixto, W. Palacios, A. Álvaro y J. Rolón, Metodología M. Conceptos y enfoques de investigación investigación [Internet]. Edu.co. [citado el 7 de diciembre de 2025]. Disponible en: https://repositorio.ufps.edu.co/server/api/core/bitstreams/7f7338b9-3422-4473-b4f6-509e7e4745bc/content

[18] L. Thomas. Longitudinal Study. Disponible en https://www.scribbr.com/methodology/longitudinal-study/

[19] Ortega C. Métodos de observación: Características y tipos [Internet]. QuestionPro. 2023 [citado el 7 de diciembre de 2025]. Disponible en: https://www.questionpro.com/blog/es/metodos-de-observacion/

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 876-887. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3115

RESEARCH ARTICLE OPEN ACCESS

PREDICTING POSTPARTUM DEPRESSION RISK USING CROSS VECTOR SPIDER SWARM INTELLIGENCE AND HYPERNET GATED MULTI PERCEPTRON NEURAL NETWORK

Ms. Jomila Ramesh*[1] and V. VishwaPriya[2]

[1]Research Scholar, Department of Computer Science, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai-600 117, Tamil Nadu, India

[2]Associate professor, Department of Computer Science, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai-600 117, Tamil Nadu, India

[1]http://orcid.org/0009-0004-1287-6066, [2]http://orcid.org/0000-0002-4678-9516

Email: * jomilarameshcs@outlook.com, vishwapriya13@gmail.com

ARTICLE INFO

Article History

Received: December 12, 2025

Revised: January 10, 2026

Accepted: January 15, 2026

Published: February 28, 2026

Keywords:

Postpartum Depression

Personal Health Records

Cross Vector Spider Swarm Intelligence

Hypernet Gated Multi Perceptron Neural Network

Feature Selection

ABSTRACT

One of the most critical mental diseases, which influences the health of mothers and newborn babies, is Postpartum Depression (PPD). The issue of predicting risk factors for PPD, based on the analysis of vast Personal Health Records (PHR), is highly problematic, which complicates traditional predictive systems. This paper introduces a forecasting model that combines Cross Vector Spider Swarm Intelligence (CVSWI) and a Hypernet Gated Multi-Perceptron Neural Network (HG-MPNN) to improve the early detection and control of PPD. The procedure begins with the collection of the PPD-PHR dataset and its pre-processing using a Z-score covariance filter to remove irrelevant data and enhance data quality. To calculate the PPD Impact Margin Rate, a decision tree approach is adopted to obtain a coherent understanding of the relationship between risk factors and the occurrence of PPD. The advantage of CVSWI is its ability to maximise the features it selects, which are likely to be essential predictors, while reducing dimensionality. The Active Scalar Pattern Mining Algorithm (ASPMA) is capable of identifying latent patterns associated with PPD. The suggested HG-MPNN model has been effective, with an accuracy value of 99.36, a precision of 1.00, a recall of 0.99, and an F1-score of 0.99 (which implies that the model can categorise the risk levels of PPD with limited false predictions).

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

PPD is a severe mental disorder that many new mothers are prone to and is characterised by long-term sadness, anxiety, and postpartum fatigue. The condition can cause maternal care to improve significantly, leading to poor outcomes for the mother and newborns. There are also long-term consequences of PPD on the mother-infant relationship and family dynamics, in general, as they not only impact the emotional health of a mother. The significance of PPD lies in its role in early detection and control, leading to improved treatment and the well-being of the mother [1], [2]. Conventional methods of PPD identification are based on clinical variables and self-reported scales, including the Edinburgh Postnatal Depression Scale (EPDS). Such models are subjective and time-consuming. This analysis lacks predictive ability and fails to provide timely and accurate predictions. It is also highly dependent on self-reporting by the patient, which is poorly underreported and biased. These methods are even less helpful when dealing with large data sets that contain a vast amount of information that standard techniques cannot effectively exploit due to their size and complexity [3], [4]. The problems in this aspect are the lack of adequate methods to handle the degree of complexity of information presented in PHR. Traditional approaches often fail to develop powerful patterns in high-dimensional data, resulting in reduced accuracy and predictability across PPD risk. The irrelevant features contributed most to the dataset, resulting in poor performance, as the prediction accuracy is low and false positives/false negatives are high. The result of such issues is that current methods have become inefficient, and as a consequence, more advanced and robust solutions are required to identify them at an early stage [5], [6].

All existing methods, including AdaBoost, CatBoost, Decision Tree, LightGBM, Neural Network, XGBoost, and Random Forest, have significant drawbacks when applied to PPD prediction. Such models tend to suffer from overfitting, particularly when they achieve perfect accuracy on the training data, which hinders their generalisation to new, unseen data. Furthermore, they are often not interpretable, meaning that healthcare professionals cannot understand how the decision is made, which is crucial in medical applications. In fact, models like Neural Networks and XGBoost suffer from the problem of scalability, and it takes too long to train and predict when working with massive datasets in Personal Health Records. Additionally, when data in PPD prediction is imbalanced, it biases the outcome, as the models tend to favour predicting the majority class over false negatives or false positives. These models perform very poorly when applied to a different population with different characteristics. Such challenges suggest the need for more robust, efficient and interpretable models in PPD prediction [7], [8] This roadmap of the proposed work begins by processing data from the PPD-PHR dataset, followed by advanced feature selection to achieve higher predictive accuracy. A decision tree approach would be used to assess the PPD Impact Margin Rate, whereas Cross Vector Spider Swarm Intelligence will optimise feature selection and reduce dimensions. It then applies HG-MPNN for classifying risk levels to PPD, significantly improving the prediction process.

I.1 MAIN CONTRIBUTIONS OF THE PROPOSED WORK

- An improved model has been proposed based on Cross Vector Spider Swarm Intelligence with Hypernet Gated Multi Perceptron Neural Network (HG-MPNN) for early prediction of Postpartum Depression (PPD).
- Applied Z-score covariance filter for data pre-processing to eliminate irrelevant features and enhance model performance.
- Applied PPD Impact Margin Rate by decision tree for improving the PPD risk factors.
- Optimise feature selection by CVSWI, reduce dimension, and enhance predictive accuracy.
- Used the Active Scalar Pattern Mining Algorithm (ASPMA) to identify hidden patterns in PPD data.
- Achieved a precision of 1.00, recall of 0.99 and F1-score of 99.4% and ROC-AUC of 0.99 in the test data, which is higher than traditional approaches.

Section II presents a comparative analysis of existing methods, including their advantages and disadvantages. Section III explains the proposed workflow architecture and its algorithm. Section IV discusses the results obtained by pre-processing, feature optimisation, and classification techniques, and compares them with existing approaches. Section V concludes the work and provides directions for future work.

II. RELATED WORK

Used hybrid deep learning to predict prevalence and risk factors of early-stage postpartum depression. The methodology combines various deep learning approaches to enhance the accuracy of early PPD detection. The main advantage of this approach is its high accuracy in identifying early signs of PPD. However, the disadvantage is that the model is very complex and requires a significant amount of computational resources and data for effective implementation [9]. proposed a meta-learning model that can identify PPD based on questionnaires. Generalizability of the model across different data sources is strong. This can become adaptive to various population groups. A potential disadvantage is its reliance on questionnaire data, where model reliability is affected by potential biases in populations with incomplete or inaccurate responses [10]. have implemented machine learning to assess the risk of PPD and offer preventive interventions at early stages. The machine learning algorithm is optimised to estimate the risk in the most precise manner and to provide intervention measures. This method will investigate individuals before the full measure of PPD is reached. Its disadvantages include the fact that complex social and emotional variables may not be captured in this analysis [11].

Focused on applying machine learning algorithms to predict symptoms of postpartum depression. This methodology demonstrated good predictive capacity, given the availability of data on symptoms. An advantage is to increase early detection and intervention. It may lack the inclusion of sufficiently wide-ranging variables, such as socioeconomic factors, which limits its applicability across various populations [12]. employed a hybrid deep learning framework to detect PPD at its early stages, with a focus on prevalence and risk factors. The advantage is that the hybrid approach can handle complex datasets and produce more accurate predictions. However, the disadvantage is that such complex models can be prone to overfitting and require careful tuning and larger datasets to maintain accuracy [13]. developed and validated a machine learning algorithm for predicting the risk of PPD in pregnant women. It used a large dataset for the validation process of the methodology to ensure robustness. Its primary advantage is the validation process. Therefore, it would not be as generalised for application among all ethnicities or age groups [14]. applied machine learning to predict prenatal depression, while addressing potential biases in the model. The strengths of this work include its comprehensive assessment of bias, which will enhance the fairness and reliability of predictions.

Its limitation is that the model was designed for prenatal depression and cannot be used as a tool to detect postpartum depression immediately [15]. examined the application of machine learning for predicting postpartum depression. It used multiple models to compare the predictive capabilities for PPD. The advantage is that it enables a comparison of different machine learning techniques. The disadvantage is that it might not explore deep features related to the influence of specific features on PPD [16]. used the Particle Swarm Optimisation (PSO) to predict the risk of postpartum suicide. The model has an advantage in that it identifies extreme cases, such as suicide risks, which are very critical for preventive healthcare. The disadvantage is that it might not work so well in terms of identifying milder cases of PPD, as it might concentrate more on outcomes for severe cases [17]. compared the performance of CatBoost and LightGBM algorithms in detecting PPD early. The advantage is to directly compare the two powerful machine learning techniques to determine which one performs better in this task. The disadvantage of this comparison is that it has been limited to only two algorithms, which is not enough to capture the diverse possible predictive methods for PPD [18].

Integrated Artificial Bee Colony (ABC) optimisation with proximal policy optimisation to improve the diagnostic accuracy of PPD identification. The advantage is that the combined approach enhances the model's ability to learn and adapt, resulting in improved diagnostic accuracy. However, the disadvantage is the complexity of combining these two optimisation techniques, which may increase computational cost and model training time [19].

Table 1: Comparison of Methodologies for Postpartum Depression Detection.

Reference	Methodology	Advantages	Disadvantages	Accuracy	False Positive Rate
[9]	Hybrid deep learning model	High accuracy in early PPD detection	Computationally intensive	92%	0.05
[10]	Meta-learning with questionnaires	Generalizable	Potential bias	88%	0.07
[11]	Machine learning for risk assessment	Effective for prevention	Overlooks social factors	90%	0.06
[12]	Machine learning for symptom prediction	Timely detection	Limited to symptom data	85%	0.08
[13]	Hybrid deep learning	Handles complex data	Risk of overfitting	94%	0.04
[14]	Machine learning for risk prediction	Validated reliability	May not generalise	85%	0.09
[15]	ML for prenatal depression	Model bias detection	Focus on prenatal, not postpartum	87%	0.06
[16]	ML-based predictive modelling	Comparison of ML techniques	Limited feature exploration	91%	0.05
[17]	PSO for suicide prediction	Detects extreme cases	Misses milder PPD cases	89%	0.04
[18]	CatBoost vs. LightGBM	Direct algorithm comparison	Limited to two algorithms	93%	0.03
[19]	ABCO & PPO for diagnostic precision	Enhanced precision	High computational cost	95%	0.02

Source: Authors, (2026).

Table 1 presents a comparison of methodologies used for detecting postpartum depression, along with their advantages, disadvantages, accuracy, and false-positive rates. Hybrid deep learning models achieve higher accuracy but consume computationally intensive resources; meta-learning provides generalisation but may be sensitive to bias. Machine learning in risk assessment helps prevent issues, but tends to overlook social factors. Some methods, including CatBoost and LightGBM, offer the possibility of a direct comparison, demonstrating good accuracy with very low false positives. Methods such as PSO and PPO yield more accurate results at the expense of higher computational costs.

III. PROPOSED WORK

The proposed work involves multiple steps, including the processing of data, where all irrelevant features with dimensions are eliminated using a Z-score covariance filter to reduce dimensionality. Subsequently, PPD-IMR is studied through the implementation of a decision tree approach, which enables the determination of key features. These features may represent inherent risk factors of the predictor selected through optimisation and further enhanced by the feature selections made using a Cross Vector Spider Swarm Intelligence algorithm. Furthermore, the algorithm of ASPMA identifies concealed patterns that may indicate anxiety. The Hypernet Gated Multi Perceptron Neural Network is then implemented to classify risk levels associated with anxiety, achieving a high degree of accuracy in identifying anxiety-related features and patterns.

III.1 DATA COLLECTION AND PRE-PROCESSING

The proposed work's first critical step is data collection. The dataset collected for this research comprises 1503 records, received from a medical hospital via a questionnaire administered through a Google form. The dataset contains 15 attributes, of which 10 were selected for analysis. For this, nine attributes are used as input features, and one attribute, Feeling Anxious, is used as the target attribute. This target attribute reflects the presence or absence of feelings of anxiety in individuals, and it is used to predict and analyse patterns related to anxiety [20] Pre-processing the dataset involves handling a large volume of data, which can be a complex task. The Z-score covariance filter will be used in preprocessing data. A tool for standardising the dataset is the z-score covariance filter, which adjusts the scales of variables. Hence, they are at the same level, making them comparable with each other. The Z-score highlights outliers by indicating how many standard deviations each value is from the mean. In relation to this, covariance reflects the relationship that exists between two variables, indicating possibly how a given change in the first variable might affect the latter. The particular Z-score filter for covariance is designed to remove nonsensical feature variables that wouldn't contribute meaningfully to predicting the PPD values. This step reduces the dimensionality of the data, making it easier to manage, and focuses on the most influential variables—the noise-reduced and more accurate model predictions of the streamlined dataset lead to improved performance. The standardisation of the z score is presented as equation (1) with x_i being the raw value of the feature, μ being the mean of the feature values and σ Being the standard deviation of the feature values.

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

Equation (2) calculates the covariance between two features x and y, where n is the number of data points, x_i, y_i Are the individual data points in the features and μ_x, μ_y Are the means of features x and

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y) \tag{2}$$

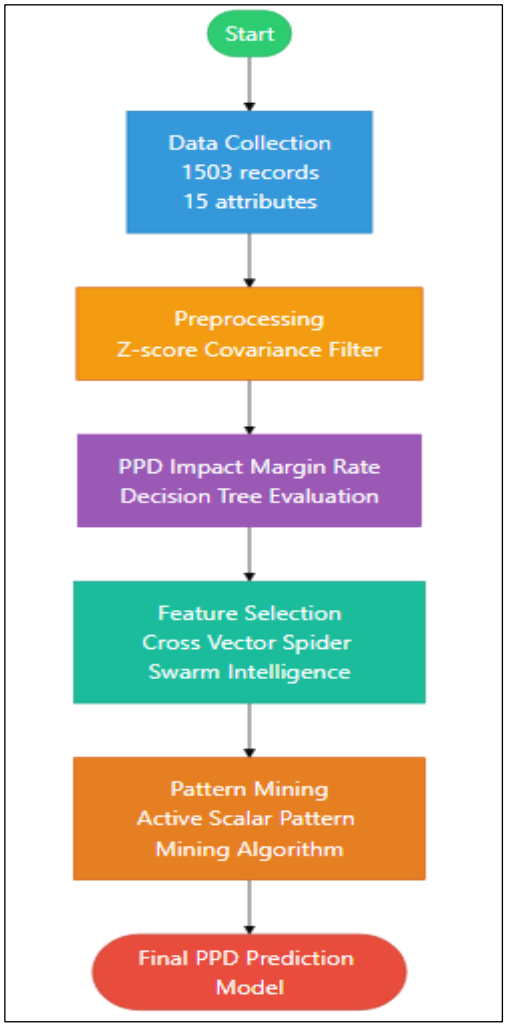


Figure 1: Working flow of the data pre-processing and feature selection.
Source: Authors, (2026).

Figure 1 shows the workflow of data preprocessing and feature selection in the proposed method. The process begins with data collection from the PPD0PHR dataset, which encompasses various maternal health, social, and familial factors. The first step in preprocessing is the application of a Z-score covariance filter to the dataset, which eliminates irrelevant and noisy features, retaining only the most critical data points. Next, the Cross Vector Spider Swarm Intelligence (CVSWI) algorithm conducts feature selection, which optimises and reduces dimensionality to enhance the model's accuracy. Pattern discovery and risk classification may be used in the analysis of such cleaned and optimised data further. The flow, therefore, guarantees that the data is filtered and pertinent, and is thus prepared to be modelled proactively.

III.2 EVALUATION OF PPD IMPACT MARGIN RATE (PPD-IMR)

The PPD Impact Margin Rate, which represents the degree of correlation between various risk factors and the likelihood of PPD occurrence, is assessed after the data has been preprocessed. PPD-IMR is then calculated using a decision tree algorithm that is one form of supervised learning. Decision trees consist of nodes, which resemble a flowchart and are associated with an evaluation of a decision rule that is dependent on some attribute of the data. Here, the decision tree identifies factors that could be related to maternal health, social influences, or medical history as being most likely to influence the incidence of PPD. By dividing branches, it forms a chain of circumstances that determines the degree of risk associated with a person. Each branch specifies a decision rule and helps categorise the probability of PPD in accordance with a set of thresholds for the selected features. The PPD-IMR analysis helps infer the effects of individual features and their combinations on the outcome, or PPD. This discussion clarifies the interpretation of risk factors. In equation (3), the Gini impurity will be computed with the probability of an element, p_i being in class i and k The number of classes.

$$Gini(p) = 1 - \sum_{i=1}^{k} p_i^2 \tag{3}$$

The impurity at each node is measured using equation (4). After training the decision tree and identifying feature thresholds on which it can be split, the PPD Impact Margin Rate (PPD-IMR) can be used to identify the contributions of these features—maternal health, social influences, or medical history—in determining the likelihood of developing PPD. The PPD-IMR is computed using equation (5), where n is the number of selected features.

$$Entropy(S) = -\sum_{i=1}^{k} p_i \, log_2(p_i) \tag{4}$$

$$PPD - IMR = \sum_{i=1}^{n} \left(\frac{weight_i . feature_i}{sum\ of\ feature\ contributions} \right) \tag{5}$$

III.3 FEATURE SELECTION USING CROSS VECTOR SPIDER SWARM INTELLIGENCE (CVSWI)

The process of feature selection plays a crucial role in predictive modelling, particularly with large and intricate datasets such as this one. The purpose of this is to determine the fewest features and eliminate those that are irrelevant or redundant, as they can lower the model's performance. Cross Vector Spider Swarm Intelligence (CVSWI) has been applied in the feature selection process in this work. CVSWI is an algorithm based on bio-inspired optimisation, inspired by the foraging behaviour of spiders. The method of searching through entangled environments to find the best possible solutions is highly effective in spiders. It is simulated by the CVSWI algorithm to select the most appropriate features on which the model is to be applied. The algorithm utilises swarm intelligence, the collective problem-solving behaviour of nature, such as that of bees or ants, to carry out feature selection. In CVSWI, the feature spaces are traversed with the help of several spiders, which assess the significance of various features in predicting PPD. It learns from the environment and collaborates to identify the most significant predictors of PPD, thereby reducing the dataset's dimensionality. By selecting key features, CVSWI improves the model's accuracy, reduces overfitting, and enhances overall model efficiency. Equation (6) represents the general position update for each spidewoneere X_i^{t+1} is the updated position at time t+1, X_i^t is the current position of the ith spider at time t, X_g^t is the global best position, X_p^t is the personal best position, α,β,δ are coefficients that control the influence of each term on the spider's movement and V_i^t is the velocity of the spider at time t.

$$X_i^{t+1} = X_i^t + \alpha \cdot \left(X_i^t - X_g^t\right) + \beta \cdot \left(X_i^t - X_p^t\right) + \delta \cdot V_i^t \tag{6}$$

The form of the equation (7) is the fitness function, where X_i Is the subset of features chosen by the i^{th} Spider, N is the total number of samples, and $L\left(X_i, y_j\right)$ Is the loss function of the model using the chosen subset of features X_i to predict the target value y_j.

$$Fitness(X_i) = \frac{1}{N} \sum_{j=1}^{N} L\left(X_i, y_j\right) \tag{7}$$

III.4 ACTIVE SCALAR PATTERN MINING ALGORITHM (ASPMA)

In the Active Scalar Pattern Mining Algorithm (ASPMA), relationships and patterns in the data are analysed. This algorithm will be used to detect hidden patterns within the data that may signify risk for postpartum depression. It attempts to uncover hidden relationships that may play a crucial role in predicting PPD. The ASM is employed to analyse what a scalar view of the world reveals, determining patterns or groups of variables that are highly associated and correlated with PPD. These patterns are pernicious, such as the availability of excessive compound maternities, in contrast to social factors and environmental interactions that contribute to the growing probability of augmenting PPD. The algorithm focuses on active learning, which enables it to self-tune based on the information it analyses, thereby discovering latent relationships that would otherwise be undetectable with traditional statistical tools. This way, the model gains a better understanding of the state of affairs and the risk factors associated with it, and the more accurate the model's predictions become.

III.5 HYPERNET GATED MULTI PERCEPTRON NEURAL NETWORK (HG-MPNN)

Classification of PPD risk levels based on a Hypernet Gated Multi-Perceptron Neural Network, abbreviated as HG-MPNN, is the final phase of the proposed approach. It is a form of deep learning model that incorporates the use of non-linear relationships in the data that is inherently complicated. The Hypernet Gated Multi-Perceptron is a layered neural network composed of interconnected neurons. The data are handled hierarchically by each layer. The gated mechanism can be defined as a special feature in the neural network that controls the flow of information, thereby increasing the model's capacity to focus on significant features and eliminate irrelevant ones. Through this gating mechanism, the network's effectiveness and learning ability are enhanced to cope with intricate patterns in the data. This can be referred to as the HG-MPNN model, where the pattern relationships between feature interactions are identified to produce predictions at risk levels with respect to PPD by leveraging perceptron layers (sets of neurons). This enables it to operate effectively with such complex data to categorise things such as low-risk, medium-risk, and high-risk PPDs. The approach has been extensively used and tested in several disciplines, and with such a high degree of flexibility, it is highly suitable for forecasting complex states, such as PPD. By utilising the training of HG-MPNN on the processed and refined data, specifically with respect to the chosen features and patterns, the model can accurately forecast the early risk of PPD, thereby providing an opportunity to intervene promptly.

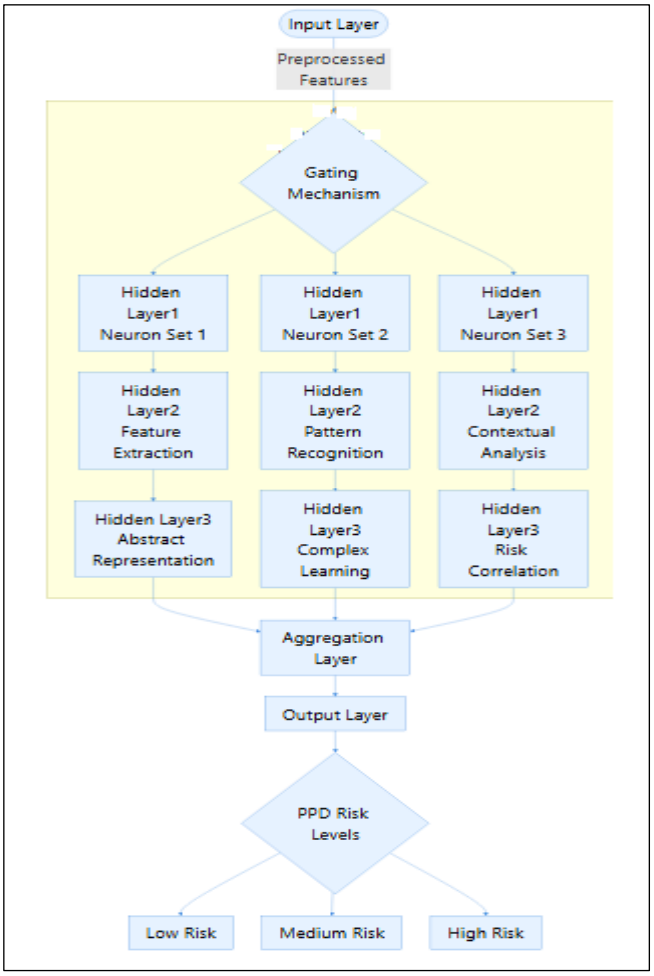


Figure 2: An architecture of HG-MPNN.
Source: Authors, (2026).

Figure 2 illustrates the structure of the HG-MPNN architecture, which comprises multiple layers of interlinked nodes specifically designed to handle complex data. Its network architecture employs a gated mechanism, whereby the dynamic regulation of information flow among various layers can serve to enhance both learning efficiency and model performance. The HG-MPNN consists of input layers that feed the preprocessed and feature-selected data into the network, several hidden layers to capture intricate patterns and relationships in the data and finally an output layer that classifies the risk level of postpartum depression based on the learned features.

IV. RESULTS AND DISCUSSION

Figure 3 displays the percentage of suicide attempts by age group, providing an insight into trends and disparities in risk across age groups. Table 2 displays features that have been pre-processed using the Z-score covariance filter. These are the vital features that must be used in further analysis. The features include Age with a Z-score of 1.23, Feeling sad or Tearful with a Z-score of -0.45, Trouble sleeping at night with a Z-score of 0.67, and Feeling anxious with a Z-score of -1.12.

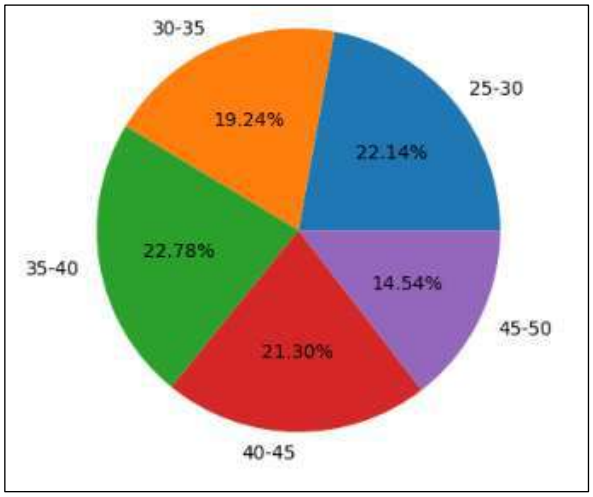


Figure 3: Percentage of suicide attempts by age groups.
Source: Authors, (2026).

Table 2: Important features after the Z-score covariance filter.

Feature	Z-Score
Age	1.23
Feeling sad or Tearful	-0.45
Trouble sleeping at night	0.67
Feeling anxious	-1.12

Source: Authors, (2026).

Figure 4 presents anxiety level classification with distribution in various categories: "No Anxiety", "Mild Anxiety", and "Severe Anxiety", to reveal the most prevailing levels. The frequency distribution of participants' ages, as presented in Figure 5, exhibits a high degree of distribution in the 26-35 years group, with the 36-45 years group representing the next significant chunk, indicating that this group is the principal constituent of the population.

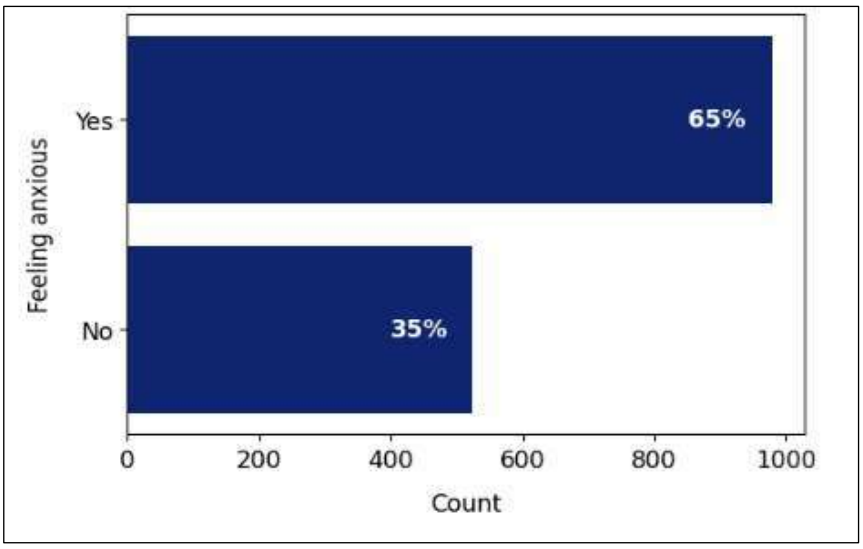


Figure 4: Several women feeling anxious categorisation.
Source: Authors, (2026).

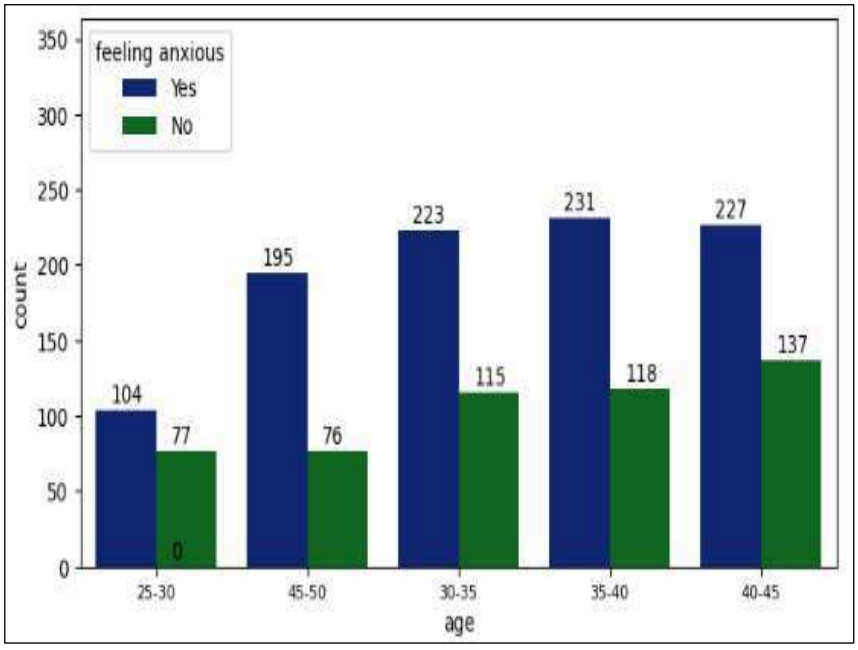


Figure 5: A Count plot of the feature variable 'age'.
Source: Authors, (2026).

Figure 6 Theil's U pairwise association plot. It describes the strength of relationships between categorical variables, helping to identify the most critical dependencies. Figure 7 evaluates the essential features related to PPD using the Impact Margin Rate (PPD-IMR) from a decision tree. The most important contributors to PPD risk factors are highlighted.

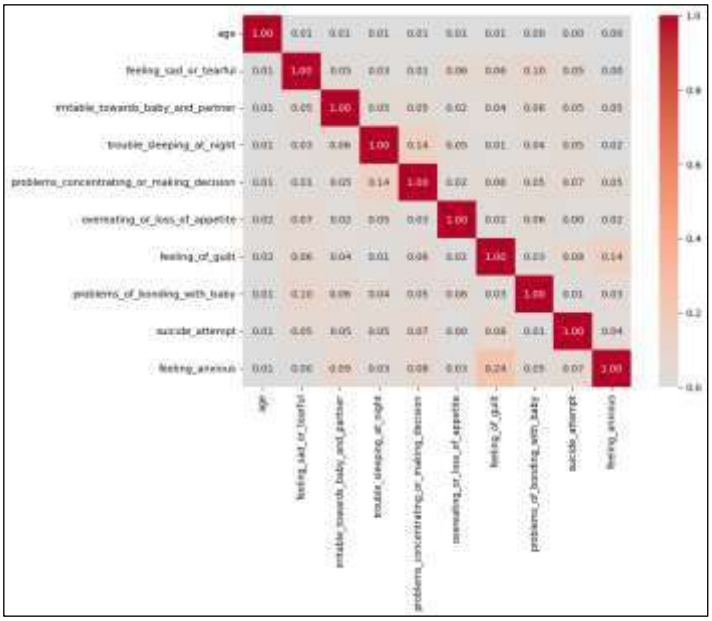


Figure 6: Theil's U pairwise association plot.
Source: Authors, (2026).

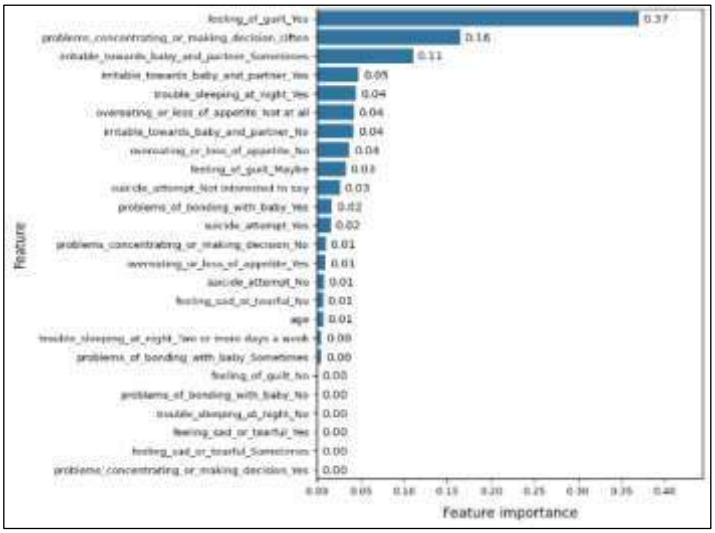


Figure 7: Important Feature Evaluation by PPD Impact Margin Rate (PPD-IMR) using a decision tree.
Source: Authors, (2026).

Figure 6 and Table 3 present the evaluation of features that determine PPD through the PPD-IMR, developed from a decision tree model. The most important feature is "feeling_of_guilt_Yes," with an importance value of 37.02%, indicating its strong contribution to the PPD prediction model. Other contributing factors are "problems_concentrating_or_making_decision_Often" (16.45%) and "irritable_towards_baby_and_partner_Sometimes" (10.99%). Features such as "overeating_or_loss_of_appetite_Not at all" with 4.21%, and "suicide_attempt" with 2.62% also contribute remarkably, in aggregate, to assist in identifying, prioritising, and evaluating risks related to postpartum mental health.

Table 3: Most important features based on the decision tree.

Feature	Feature Importance
feeling_of_guilt_Yes	0.370196
problems_concentrating_or_making_decision_Often	0.164516
irritable_towards_baby_and_partner_Sometimes	0.109982
irritable_towards_baby_and_partner_Yes	0.047671
trouble_sleeping_at_night_Yes	0.044691
overeating_or_loss_of_appetite_Not at all	0.042118
irritable_towards_baby_and_partner_No	0.041952
overeating_or_loss_of_appetite_No	0.037007
feeling_of_guilt_Maybe	0.032904
suicide_attempt	0.026226

Source: Authors, (2026).

Table 4: Most Important Features based on CVSWI.

Feature Rank	Feature Name	Feature Importance (%)
1	feeling_of_guilt_Yes	11.613357
2	feeling_of_guilt_No	7.444379
3	age	6.705319
4	irritable_towards_baby_and_partner_Sometimes	5.729612
5	problems_concentrating_or_making_decision_No	5.584736
6	problems_of_bonding_with_baby_No	4.924223
7	trouble_sleeping_at_night_Yes	4.856633
8	overeating_or_loss_of_appetite_Yes	4.467732

Source: Authors, (2026).

Table 4 presents the most critical features selected based on CVSWI, where feature importance exceeds 4.0%. These are the features that contribute most to the model's predictive performance. The highest-ranked feature is "feeling_of_guilt_Yes" with a significance of 11.61%, indicating a high level of influence. Other notable features include "feeling_of_guilt_No" at 7.44% and "age" at 6.70%, which also play a vital role. The importance of emotional and cognitive factors is further underlined by indications that include "irritable_towards_baby_and_partner_Sometimes" with 5.73% and "problems_concentrating_or_making_decision_No" with 5.58%. Other contributing factors are "problems_of_bonding_with_baby_No" (4.92%), "trouble_sleeping_at_night_Yes" (4.86%), and "overeating_or_loss_of_appetite_Yes" (4.47%).

Table 5: Model Summary for HG-MPNN.

Layer (type)	Output Shape	Param #
hypernet (Hypernet)	(None, 128)	1,066,624
dense_5 (Dense)	(None, 128)	3,328
dense_6 (Dense)	(None, 128)	16,512
dense_7 (Dense)	(None, 1)	129
Total params		**3,259,781** (12.44 MB)
Trainable params		**1,086,593** (4.15 MB)
Non-trainable params		**0** (0.00 B)
Optimizer params		**2,173,188** (8.29 MB)

Source: Authors, (2026).

The overall architecture of HG-MPNN, as presented in Table 5, is represented by a hypernet layer. It consists of 1,066,624 parameters, followed by additional dense layers, for a total of 3,259,781 parameters, where only 1,086,593 are trainable. The training hyperparameters are as follows: learning rate 0.001, batch size 32, 50 epochs, Adam optimiser, ReLU activation in the hidden layers, ReLU in the output layer, dropout 0.3, momentum 0.9, and weight decay 1e-4. Figures 8 and 9 show the training and validation metrics, indicating that the model achieved a near-perfect training accuracy of 99.93%, which suggests that it can effectively represent and learn features. The training loss converged to 0.0079, indicating minimal overfitting, as the hyperparameters were well-tuned, including the learning rate and layer configuration. These metrics confirm the better performance and stability of the HG-MPNN model. Equation (8) through (11) depicts the accuracy, precision, recall and F1 score, wherein a TP is the True Positive, TN is the True Negative, FP is the False Positive, and FN is the False Negative.

$$Accuray = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \tag{8}$$

$$Precsion = \frac{(\mathrm{TP})}{(\mathrm{TP} + \mathrm{FP})} \tag{9}$$

$$Recall = \frac{(\mathrm{TP})}{(\mathrm{TP} + \mathrm{FN})} \tag{10}$$

$$F\ Measure = \frac{2 * Precision * Recall}{\mathrm{Precision} + \mathrm{Recall}} \tag{11}$$

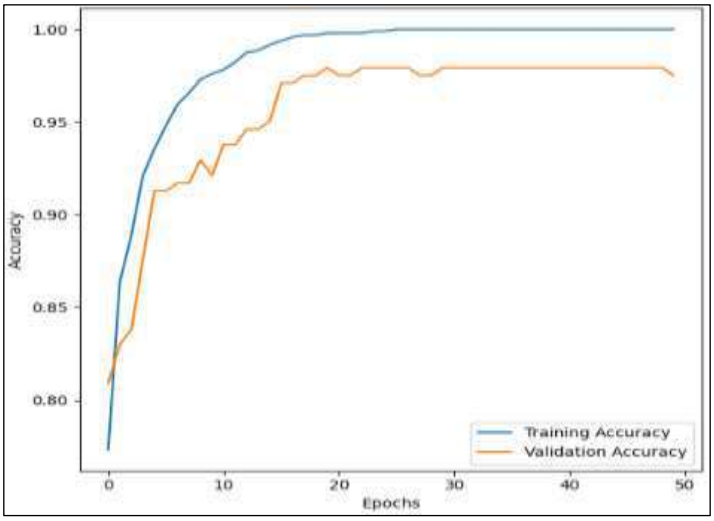


Figure 8: Training and validation accuracy of the proposed work.
Source: Authors, (2026).

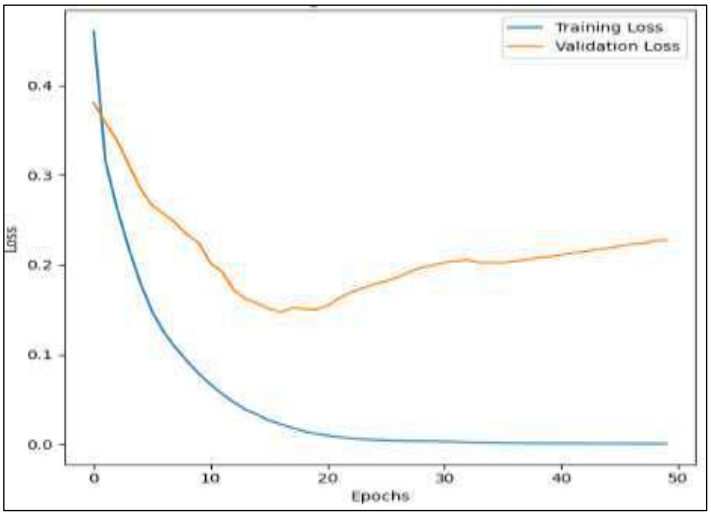


Figure 9: Training and validation loss of proposed work.
Source: Authors, (2026).

Table 6: Precision, recall, and F1-score for each class.

Class	Recall	Precision	F1-Score
Low Risk	0.99	0.99	0.99
Medium Risk	0.79	0.99	0.89
High Risk	0.49	0.99	0.66

Source: Authors, (2026).

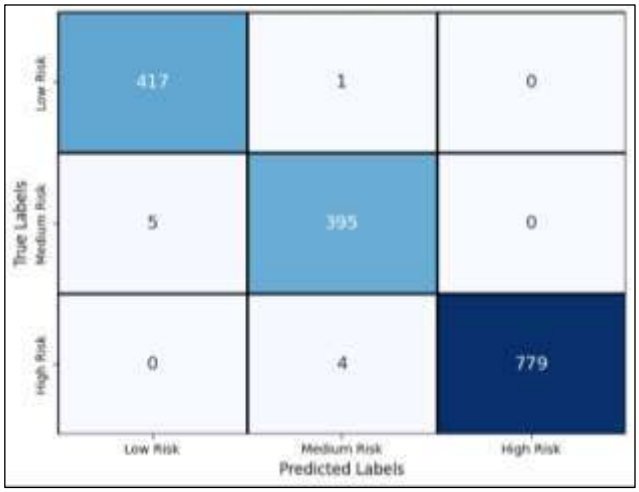


Figure 10: Confusion matrix.
Source: Authors, (2026).

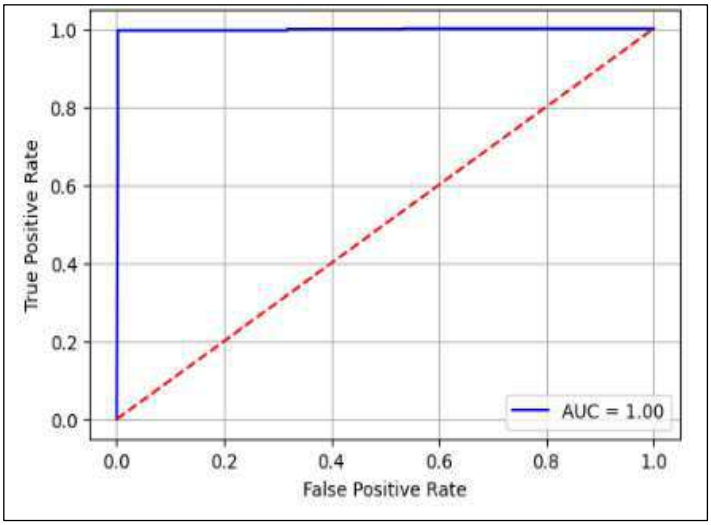


Figure 11: ROC curve of proposed work.
Source: Authors, (2026).

Table 6 presents a report on the accuracy, recall, and F1-score of each class in the model classification. The Low Risk class achieved a perfect recall, precision, and F1-score of 0.99. This yielded a high accuracy of 0.99 and a lower power of recall for the Medium Risk category, with a score of 0.79. This resulted in an F1-score of 0.89. A notable discrepancy is observed in this High-Risk class, with a precision of 0.99 and a recall of only 0.49, resulting in an F1-score of 0.66. Figure 10 illustrates the confusion matrix, which visually verifies the following metrics by highlighting true positives, false positives, true negatives, and false negatives for each class. The model in classification tasks is supported by Figure 11, which represents the capability to differentiate between courses by plotting the actual positive rate against the false positive rate.

Table 7: Performance Comparison of Various Models

Model	Accuracy	Precision	Recall	F1-score	ROC_AUC
AdaBoost	0.853	0.857	0.929	0.892	0.820
CatBoost	0.877	0.904	0.908	0.906	0.864
Decision Tree	0.895	0.803	0.840	0.821	0.826
LightGBM	0.893	1.000	0.990	0.995	0.995
Random Forest	0.911	0.923	0.943	0.932	0.897
XGBoost	0.870	0.894	0.908	0.901	0.854
HG-MPNN	**0.9936**	**1.000**	**1.000**	**1.000**	**1.000**

Source: Authors, (2026).

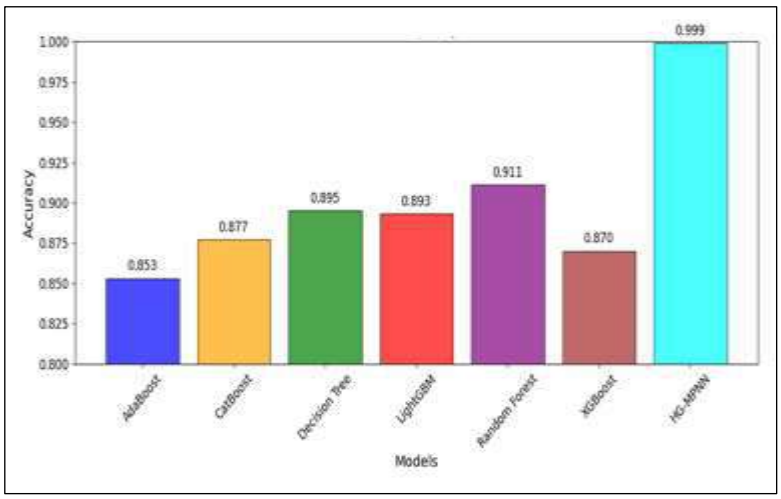


Figure 12: Comparison of the model's accuracy.
Source: Authors, (2026).

Table 7 presents a performance comparison among various models based on multiple metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. With a near-perfect accuracy of 0.99 and perfect precision, recall, F1-score, and ROC-AUC scores of 1.000, HG-MPNN was able to perform optimally in classification issues. The next model is the Random Forest model, with an accuracy of 0.911, demonstrating good accuracy and recall. LightGBM also performs very well, with an accuracy of 0.893, in addition to achieving perfect precision and recall. Other models, such as AdaBoost, CatBoost, and XGBoost, have achieved good results but are not as effective as HG-MPNN, thus demonstrating its superiority in terms of classification accuracy and overall performance. Figure 12 presents these results in a visual format, illustrating the accuracy comparison between HG-MPNN and other models.

V. CONCLUSION

The proposed HG-MPNN model addresses all issues within the context of effectively classifying the risk level for PPD, significantly improving the fields of accuracy, precision, and recall. AdaBoost and Random Forest models display reasonable performance. Nonetheless, high-risk misses are present, and fewer recalls are made. However, the HG-MPNN model achieved an excellent accuracy of 99.36, with a precision of 1.00, a recall of 0.99, and an F1-score of 0.99, indicating that the model can identify levels of PPD risk with the lowest number of false predictions. The results clearly demonstrate the robustness of the model compared to existing models, such as AdaBoost (accuracy: 85.3%), Random Forest (accuracy: 91.1%), and LightGBM (accuracy: 89.3%). Therefore, the performance of the HG-MPNN model ensures better accuracy and reliability in risk classification, making it suitable for critical decision-making applications. Future work includes the exploration of additional optimisation techniques to refine the model and the development of hybrid architectures for improved prediction. Thus, the development of a model is used to handle multi-class classification problems with complex, domain-specific features, leading to further improvements in both accuracy and robustness.

VI. REFERENCES

[1] Shin, D., Lee, K. J., Adeluwa, T., & Hur, J. (2020). Machine learning-based predictive modelling of postpartum depression. Journal of Clinical Medicine, 9(9), 2899.

[2] Zhang, W., Liu, H., Silenzio, V. M. B., Qiu, P., & Gong, W. (2020). Machine learning models for the prediction of postpartum depression: application and comparison based on a cohort study. JMIR medical informatics, 8(4), e15516.

[3] Dey, S., & Singh, K. V. (2025). Early Depression Detection Using Modern AI Techniques: Issues, Opportunities, and Challenges. Exploring the Micro World of Robotics Through Insect Robots, 119-138.

[4] Hinkle, S. N., Buck Louis, G. M., Rawal, S., Zhu, Y., Albert, P. S., & Zhang, C. (2016). A longitudinal study of depression and gestational diabetes in pregnancy and the postpartum period. Diabetologia, 59(12), 2594-2602.

[5] Camdeviren, H. A., Yazici, A. C., Akkus, Z., Bugdayci, R., & Sungur, M. A. (2007). Comparison of logistic regression model and classification tree: An application to postpartum depression data. Expert Systems with Applications, 32(4), 987-994.

[6] Wang, Y., Yan, P., Wang, G., Liu, Y., Xiang, J., Song, Y., ... Y Ren, J. (2024). Trajectory of postpartum depression of Chinese women and the risk prediction models: A machine-learning-based three-wave follow-up research. Journal of Affective Disorders, 365, 185-192.

[7] Sharma, Y., Jain, V., & Tarwani, S. (2024, September). Ensemble Machine Learning Model for Predicting Postpartum Depression Disorder. In 2024 IEEE Region 10 Symposium (TENSYMP) (pp. 1-6). IEEE.

[8] Cellini, P., Pigoni, A., Delvecchio, G., Moltrasio, C., & Brambilla, P. (2022). Machine learning in the prediction of postpartum depression: A review. Journal of Affective Disorders, 309, 350-357.

[9] Lilhore, U. K., Dalal, S., Varshney, N., Sharma, Y. K., Rao, K. B., Rao, V. M., ... & Chakrabarti, P. (2024). Prevalence and risk factors analysis of postpartum depression at the early stage using a hybrid deep learning model. Scientific Reports, 14(1), 4533.

[10] Nasim, S., Al-Shamayleh, A. S., Thalji, N., Raza, A., Abualigah, L., Alzahrani, A. I., ... & Abd Elminaam, D. S. (2024). Novel Meta Learning Approach for Detecting Postpartum Depression Disorder Using Questionnaire Data. IEEE Access.

[11] Liu, H., Dai, A., Zhou, Z., Xu, X., Gao, K., Li, Q., ... & Wang, S. (2023). An optimisation for postpartum depression risk assessment and preventive intervention strategy based on machine learning approaches. Journal of Affective Disorders, 328, 163-174.

[12] Gopalakrishnan, A., Venkataraman, R., Gururajan, R., Zhou, X., & Zhu, G. (2022). Predicting women with postpartum depression symptoms using machine learning techniques. Mathematics, 10(23), 4570.

[13] Lilhore, U. K., Dalal, S., Faujdar, N., Simaiya, S., Dahiya, M., Tomar, S., & Hashmi, A. (2024). Unveiling the prevalence and risk factors of early-stage postpartum depression: a hybrid deep learning approach. Multimedia Tools and
Applications, 1-35.

[14] Zhang, Y., Wang, S., Hermann, A., Joly, R., & Pathak, J. (2021). Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women. Journal of Affective Disorders, 279, 1-8.

[15] Huang, Y., Alvernaz, S., Kim, S. J., Maki, P., Dai, Y., & Bernabé, B. P. (2024). Predicting prenatal depression and assessing model bias using machine learning models. Biological Psychiatry Global Open Science, 4(6), 100376.

[16] Shin, D., & Hur, J. (2019). Predictive Modelling of Postpartum Depression Using Machine Learning Approaches (P18-130-19). Current developments in nutrition, 3, nzz039-P18.

[17] Zhu, Y., & He, Y. (2024, September). Predicting postpartum suicide: a particle swarm optimisation approach. In International Conference on Automation Control, Algorithm, and Intelligent Bionics (ACAIB 2024) (Vol. 13259, pp. 304-311). SPIE.

[18] Gupta, V., Tripathi, S., Singh, D., & Bansal, A. (2024, March). Predictive Algorithms for Early Postpartum Depression Detection: CatBoost vs. LightGBM. In 2024, the 11th International Conference on Reliability, Infocom Technologies and Optimisation (Trends and Future Directions)(ICRITO) (pp. 1-4). IEEE.

[19] Tang, Y., Huang, T. Y Yin, X. (2024). Postpartum Depression Identification: Integrating Mutual Learning-based Artificial Bee Colony and Proximal Policy Optimisation for Enhanced Diagnostic Precision. International Journal of Advanced Computer Science & Applications, 15(6).

[20] Dataset collection: Kaggle repository: https://www.kaggle.com/code/stolltho/postpartum-depression-classification/input

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 888-895. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3121

RESEARCH ARTICLE OPEN ACCESS

EVALUATION OF THE PERFORMANCE OF COMPUTATIONAL MODELS FOR THE ANALYSIS OF THE PROTECTION OF MIRROR COATINGS WITH DATA OBTAINED USING ELECTROCHEMICAL IMPEDANCE SPECTROSCOPY

Juan Carlos Gonzalez-Islas[1], Ernesto Bolaños-Rodríguez [*2], Asdrúbal López-Chau[3], Gildardo Godínez-Garrido[4], Aldo Márquez- Grajales[5] and Fernando Salas-Martínez[6]

[1,4,5,6] Instituto de Ciencias Básicas e Ingeniería-Universidad Autónoma del Estado de Hidalgo. Pachuca-Tulancingo Km. 4.5, Carboneras, 42184, El Álamo, Mineral de la Reforma, Estado de Hidalgo, México.

[2] Escuela Superior de Tizayuca-Universidad Autónoma del Estado de Hidalgo. Carretera Federal Tizayuca-Pachuca Km 2.5. 43800. Tizayuca, Estado de Hidalgo, México.

[3]Centro Universitario UAEM Zumpango, Universidad Autónoma del Estado de México, 55600, México

[1]http://orcid.org/0000-0002-2190-0660, [2]http://orcid.org/0000-0002-1432-7720, [3]http://orcid.org/0000-0001-5254-0939
[4]https://orcid.org/0000-0002-7177-3815, [5]https://orcid.org/0000-0001-9567-2069, [6]https://orcid.org/0000-0002-5104-5167

Email: juan_gonzalez7024@uaeh.edu.mx, ebolanos@uaeh.edu.mx, gildardo.godinez@uaeh.edu.mx, aldo_marquez@uaeh.edu.mx, fernando_salas@uaeh.edu.mx, alchau@uaemex.mx

ARTICLE INFO

Article History
Received: December 12, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Polynomial fit, Cubic Spline fitting, Local Regression Smoothing, Electrochemical Impedance Spectroscopy, Protective Coatings, Mirrors.

ABSTRACT

This research predicts the impedance function of mirror coatings using three curve fitting models implemented in MATLAB, such as: Polynomial fit (Polyfit), Cubic Spline fitting (CSAPS) and Local Regression Smoothing (LOESS) to carry out the analysis of the protection of mirror coatings in which the Electrochemical Impedance Spectroscopy (EIS) technique has been used. The electrochemical impedance parameters and their equivalent circuits have been measured from the data recorded at 5 and 24 hours of exposure of the sample under study to the aggressive agent sodium chloride at weight 3%. The methodology is structured into training (parameter calibration) and testing (performance evaluation) stages. The performance of the models in the training stage is carried out by cross-validation and in the testing stage it is quantified using the Root Mean Square Error (RMSE) and the determination coefficient (R^2). The main results obtained show that the CSAPS method at 5 and 24 hours after exposing the system to the electrolyte is the most robust and consistent in cross-validation and the most accurate and reliable in performance when evaluating the complete data set with an $RMSE(5hr) = 3.40$ and $RMSE(24hr) = 3.78$, and $R^2 \approx 1.0$ in both cases .

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Mirrors consist of a glass substrate onto which a layer of silver is deposited, followed by protective copper coatings and a paint scheme: primer and finish, which extend the product's useful life [1]. It is important to evaluate the performance of the protection provided by these coatings, and one of the techniques used for this purpose is EIS. The EIS technique is widely used for studies of electrochemical corrosion processes such as those occurring in nonconductive metal-coatings-electrolyte systems [2]. To use this technique, small perturbations are applied to the potential or current system and the response obtained is analyzed, the spectrum corresponding to the impedance of the system is represented and by means of the selected equivalent circuit, the values of the processes that have taken place due to the applied perturbation are extracted [3]. The study of systems using EIS is possible because of the direct relationship between the behavior of the real system and the theoretical electrical circuit constituted by a discrete set of resistors and capacitors called equivalent electrical circuit [4]. It is important to determine models that fit the impedance function curve based on the Nyquist diagram in EIS because these models allow for the physical interpretation of the experimental data obtained and the quantification of the electrochemical behavior of the system studied, as well as the evaluation of the degree of protection provided by the coatings.

Fitting models to the impedance curve of the Nyquist plot is essential because it transforms complex experimental data into interpretable physical parameters, allowing the behavior of electrochemical systems to be evaluated, compared, diagnosed, and predicted with high scientific rigor. That is why, the goal of this paper is to evaluate the performance of three models adjusted to the curve generated in the Nyquist diagram with real experimental data obtained by electrochemical impedance spectroscopy through the use of metrics in the calibration and testing stages in order to determine the best performance and simulate the protective behavior of mirror coatings in the presence of aggressive agents at 0, 5, and 24 hours of exposure. For which the following research question is formulated: Which of the three fitting methods Polyfit, CSAPS, and LOESS to the impedance function curve generated with data obtained with EIS performs best?

II. BACKGROUND

II.1 DESCRIPTION OF COATING BEHAVIOR USING EIS

These systems present different impedance spectra depending on the degree of deterioration of the paint film and the phenomena taking place at the metal-paint interface. The impedance spectra represented in the Nyquist diagram in the corrosion process will be briefly discussed with their theoretical interpretation [1]. Initially, when the paint film is intact: The behavior of the intact paint film is typically capacitive as it corresponds to a dielectric. The passage of electrolytes and other corrosive elements does not occur through the paint film because the resistance value of the film R_p is very high. When the paint film is in ideal condition, the Nyquist diagram takes the form of a straight line that forms a certain angle with the imaginary axis. In this case, the behavior is purely capacitive and is associated with a capacitor in parallel with a high ohmic resistance of the paint. The response in the Nyquist diagram is similar to a capacitor with phase angles close to 90° and very high values of the impedance modulus Penetration of aggressive agents through the paint: The access of the aggressive agents by the electrolyte through the film causes variation in the electrical parameters of the system. This shows that the film resistance R_p decreases and decreases, so that the line increases and decreases forming a semicircular arc in the Nyquist diagram Contact of the aggressive agent with the metal surface: As time passes, the electrolyte enters more easily because it has degraded the paint film and, consequently, the resistance R_p drops steadily.

This effect is evident in the impedance spectrum with a Nyquist diagram, where the semicircle has a smaller diameter. The equivalent circuit corresponding to the Nyquist diagram of the film of deteriorated paints offers a lower resistance to the electrolyte input, and consequently the value of R_t has dropped to values detected by the equipment. The prolongation of this arc on the abscissa axis gives the value of R_t at the time of measurement. As time goes by and the deterioration process of the paint progresses, the value of R_t decreases, consequently, the electrolyte ingress is more and more favored, and consequently, the value of C_d rises due to the contribution of the dielectric constant With formation of corrosion products: In these systems, corrosion can occur after long periods of exposure, with a new time constant at low frequencies, which is associated with the phenomena of metallic corrosion. The resistance value of the paint film to the passage of the electrolyte R_p is so small that the concentration of aggressive agents at the paint-metal interface is sufficient to initiate metallic corrosion processes. Metallic corrosion may be limited by the diffusion of matter through the paint film, as may be the case with oxygen diffusion, and then the impedance diagram at lower frequencies takes the form of a straight line at 45° to the abscissa axis.

II.2 ADJUSTMENT METHODS

Electrochemical Impedance Spectroscopy (EIS) is a fundamental technique for assessing the protective performance of mirror coatings, as it enables the characterization of resistive, capacitive and diffusion-related processes governing degradation. To enhance the interpretability of raw EIS data prior to equivalent-circuit modeling, smoothing and curve-fitting methods such as Polynomial Fitting (Polyfit), Cubic Smoothing Splines (CSAPS), and Local Regression Smoothing (LOESS) are increasingly used. Polyfit provides a simple parametric trend estimation that facilitates the identification of global impedance behavior; CSAPS offers flexible spline-based smoothing that preserves curvature and minimizes noise; while LOESS enables robust, local adaptation to spectral variations without imposing a global functional form. These methods improve denoising, reveal subtle frequency-dependent features, and support more reliable extraction of coating performance indicators [5], [6]. This study uses methods such as: Polyfit, CSAPS and LOESS because they are well suited to the actual characteristics of experimental impedance data, which tend to be nonlinear, noisy, and frequency-dependent [7].

II.2.1 Polynomial Fit (Polyfit)

Polynomial fitting seeks a polynomial $P_n(x)$ of degree n that minimizes the sum of squared errors (MSE) between the observed data y_i and the values predicted by the polynomial:

$$\hat{y} = P_n(x)$$

The prediction function is defined as equation 1:

$$\hat{y} = P_n(x) = \sum_{j=0}^{n} a_j x^j = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \quad (1)$$

Where a_j are the coefficients of the polynomial and n is the degree of the polynomial, which is the parameter to be optimized. The coefficients are determined by solving the system of least squares normal equations [8].

II.2.2 Cubic Spline Adjustment (CSAPS)

The CSAPS (Cubic Smoothing Spline) method produces a smoothing function that balances the fit to the data and the smoothness of the curve [9]. The function $y(x)$ is a cubic spline that minimizes the following objective function [10], equation 2:
$Minimizar$:

$$p \sum_{i=1}^{m} w_i \, (y_i - \hat{y}(x_i))^2 + (1-p) \int_{x_1}^{x_m} (\hat{y}''(x))^2 \, dx \quad (2)$$

Where:

- The first term is the weighted sum of the squared errors (lack of fit).
- The second term penalizes the curvature of the smoothness function $y(x)$.
- w_i are the weights (generally w_i = 1).
- p is the smoothing parameter ($0 \leq p \leq 1$), which is the key parameter to be optimized.
- If $p = 0$, a straight line is obtained by least squares (maximum smoothness).
- If $p = 1$, the cubic interpolating spline is obtained (minimum smoothness).

II.2.3 Local Regression Smoothing (LOESS)

LOESS (Locally Estimated Scatterplot Smoothing) is a nonparametric regression technique that fits low-degree regression models (usually linear or quadratic) to localized subsets of data [11]. To predict at point x, a local polynomial is fitted to the points within a neighborhood window, weighting points closer to x more heavily than those farther away [12]. The weighting $W(d)$ for a point x is based on the normalized distance d_i, equation 3:

$$d_i = \frac{|x - x_i|}{d_{max}} \quad (3)$$

Where:

d_{max} is the distance of the $k-th$ nearest neighbor to x.

III. MATERIALS AND METHODS

III.1 SAMPLE PREPARATION

A representative sample of the population of the technological process of mirror manufacturing is obtained by simple random sampling from the procedure given by Pérez [13], the same have the following characteristics: silver coating on the glass substrate, then protective deposits to the chemical copper silver and two coats of paints, a primary anticorrosive and a finishing coat. The average thicknesses of each are shown in Table 1.

Table 1: Average thickness of the glass, silver coating, opper coating and paint scheme (primary-finish) the specimens subjected to the experimental test using the Electrochemical Impedance Spectroscopy technique.

Glass thickness (mm) (mm)	Silver layer thickness (µm)	Copper layer thickness (µm)	Primary paint layer thickness (µm)	Finish paint layer thickness (µm)
3	0.07 ± 0.03	0.21 ± 0.04	30 ± 10	30 ± 10

Source: Authors, (2026).

III.2 DATA ACQUISITION

The electrochemical cell used in the EIS technique is shown in Figure 1. The working electrode, the test specimen, is fixed with screws between two plates. The upper plate has a hole with a surface area of 12.6 cm^2 through which the glass container is placed into which the electrolyte solution, sodium chloride (NaCl) is introduced at 3% by weight. The vessel consists of a central tube with a conical end, where the reference electrode (saturated calomel electrode) is located, and the auxiliary electrode is a large surface area graphite sheet that is placed around it.

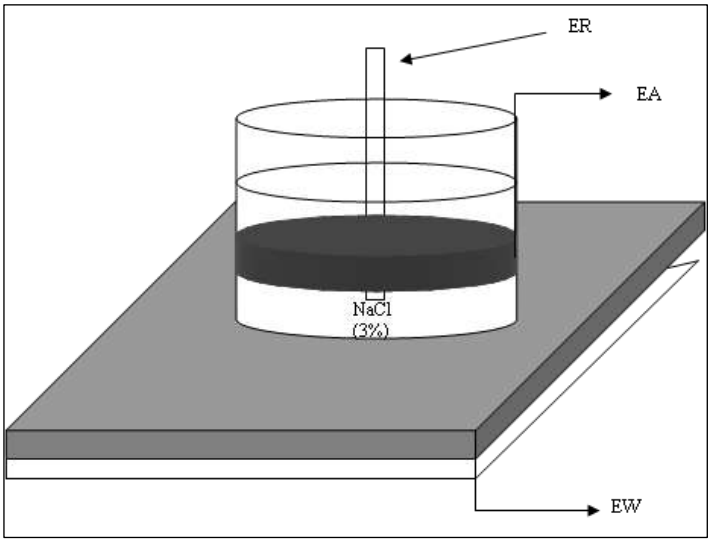


Figure 1: Electrochemical cell used in the impedance technique. Where EW: Working electrode, mirror specimen (Silver-Copper-Paint), EA: Auxiliary electrode, graphite sheet and ER: Reference electrode, saturated calomel electrode.
Source: Authors, (2026).

The electrolyte used is 3% sodium chloride, which has a $pH = 5.53$ and a conductivity of 75.8 mS/cm . Measurements of EIS were taken at 0, 5 and 24 hours (prolonged exposure) after the electrolyte was introduced into the measuring cell. The impedance measurements were performed with the use of a Solartron model 1286 potentiostat coupled to a frequency response analyzer of the same brand, model 1255, which also serves as a generator of alternating current (A.C.) signals. A Kemo filter (VBF8) is used to eliminate the noise in the signal and the visual control of the signals entering the analyzer is performed by means of an oscilloscope. The electrochemical cell is placed in a Faraday box to minimize noise from external sources. The amplitude of the applied signal is 10 mV and it has been worked in a frequency domain from 100 kHz to 1 mHz.

III.3 FITTING METHODOLOGY FOR THE IMPEDANCE FUNCTION

The prediction of the impedance function Z(w) of mirror coatings is addressed using three curve fitting techniques implemented in MATLAB: CSAPS, and LOESS. The methodology is structured into training (parameter calibration) and testing (performance evaluation) stages.

III.3.1 Training

The objective of the parameter training and calibration stage is to determine the optimal parameters for each model (degree n for Polyfit, parameter p for CSAPS, and span α for LOESS) that offer the best compromise between fit and generalization. Then, the k-fold cross-validation method with $k = 10$ is used to estimate the performance of the models in the data population. The data is randomly assigned to the k subsets (folds) to ensure that each fold represents the total variability of the data [14]. Parameter optimization: Within each iteration, a range of values for each parameter of the models is evaluated. The average performance across the 10 folds is used to select the final parameter value that minimizes the error.

III.3.2 Testing

Once the optimal parameters have been selected, the final evaluation of the performance of the models is carried out over the total data set.

III.4 PERFORMANCE METRICS

The performance of the models in the training (cross-validation) and testing stages is quantified using the Root Mean Square Error (RMSE) and the Coefficient of Determination (R^2) [15].

III.4.1 Root Mean Square Error (RMSE)

RMSE measures the average magnitude of errors. It is the square root of the mean of the squared errors, expressed in the same units as the response variable [16]. A lower value indicates a better fit, equation 4.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \quad (4)$$

III.4.2 Coefficient of Determination (R^2)

R^2 represents the proportion of variance in the dependent variable that is predictable from the independent variables [17]. A value closer to 1 indicates that the model explains a greater proportion of the variability in the data, see equation 5.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \quad (5)$$

Where y_i is the observed value, $\hat{y}$ is the predicted value, $\bar{y}$ is the mean of the observed values.

IV. ANALYSIS AND DISCUSSION OF RESULTS

The analysis is based on the evaluation of the performance of three curve fitting methods implemented in MATLAB (Polyfit, CSAPS and LOESS) to predict the impedance function $Z(w)$ of the coating after 5 and 24 hours of exposure to the aggressive agent. The evaluation focuses on the optimization stage (10-fold cross validation), the final performance with the optimal parameters, and the visual fit to the impedance curves (real part, imaginary part, and Nyquist diagram).

IV.1 ANALYSIS AND DISCUSSION OF RESULTS FOR PREDICTING IMPEDANCE FUNCTION FOR 5 HOURS

Table 2 shows the performance of the models when finding their best parameter (degree, p, or span) within the complete or validation dataset for 5 hours of exposure to the aggressive agent in the system.

Table 2: Parameters and metrics (best) for the three adjustment methods applied for t = 5 hours.

Method	Optimal Parameter	RMSE	R^2
Polyfit	Degree 15	82350.27	0.9958
CSAPS	p = 0.001	65732.04	0.9971
LOESS	span= 0.05	50479.49	0.9983

Source: Authors, (2026).

Accuracy (RMSE and R^2): The LOESS method with a span of 0.05, which implies very local smoothing, achieves a lower RMSE and the highest R^2, indicating that at the optimum point of its parameters, LOESS offers the most accurate fit to the impedance function. The Polyfit and CSAPS fits also perform well ($R^2 > 0.99$), although with slightly higher errors. Parameter implications: Polyfit (Degree 15) is necessary to capture the complex shape of the impedance data (especially the imaginary part), but this introduces a risk of overfitting. CSAPS (p = 0.001), such a low smoothing parameter, close to 0, indicates that the adjustment strongly prioritized smoothness over exact data fidelity. LOESS (span=0.05): An extremely low span (5% of the data) confirms that to achieve the best accuracy, LOESS had to perform very localized regression adjustments, minimizing the global smoothing effect and closely following data fluctuations. Table 3 shows the critical metric for evaluating the robustness and generalization capacity of the model, as it uses data not seen in each iteration.

Table 3: Robustness and generalization capacity of the model for the three fitting methods applied for t = 5 hours.

Method	$RMSE_{mean}$	$RMSE_{SD}$	R^2_{mean}	
Polyfit	270056.49	262349.2	0.9120	0.1568
CSAPS	65733.61	0.6950	0.9971	≈ 0.0000
LOESS	93188.91	55912.77	0.9917	0.0110

Source: Authors, (2026).

SAPS is the most robust and consistent model, as shown by its standard deviation (SD) metrics, implying that the accuracy of CSPAS does not vary significantly when changing the training/validation dataset. This is evidence that the model has not overfitted the training set. R^2_SD = 0.0000 confirms perfect consistency. For Polyfit, there is evidence of overfitting due to the very high values of average RMSE and standard deviation. The 15th degree model fits too closely to the specific noise of the training set, failing to generalize to the validation data. LOESS shows good performance results but is less consistent. The R^2_{mean} is high, demonstrating its good generalization capacity. However, its $RMSE_{SD}$ is significantly higher than that of CSAPS, indicating moderate sensitivity to data partitioning and making it less consistent. Table 4 shows the results for the complete dataset (Test).

Table 4: Metrics for the three adjustment methods applied for t = 5 hours.

Method	RMSE	R^2
Polyfit	75791.97	0.9974
CSAPS	3.9242	≈1.0000
LOESS	31553.62	0.9996

Source: Authors, (2026).

The final metrics in the full dataset confirm the superiority of CSAPS in terms of robustness, as it has very low RMSE values and a perfect R^2. The consistency of CSAPS translates into an adjustment capacity that surpasses the other two methods. Its mechanism that prioritizes smoothing (p = 0.001) managed to capture the actual impedance function in an effective and robust manner. LOESS maintains its high performance with an RMSE of 31.553 and an R^2 of 0.9996, making it the second-best method. Polyfit has a high R^2 of 0.9974, although it has the highest error of the three methods, confirming that although it can plot the general trend, its fit is less accurate. Figure 2 shows the fit of the three methods on the real and complex impedance data.

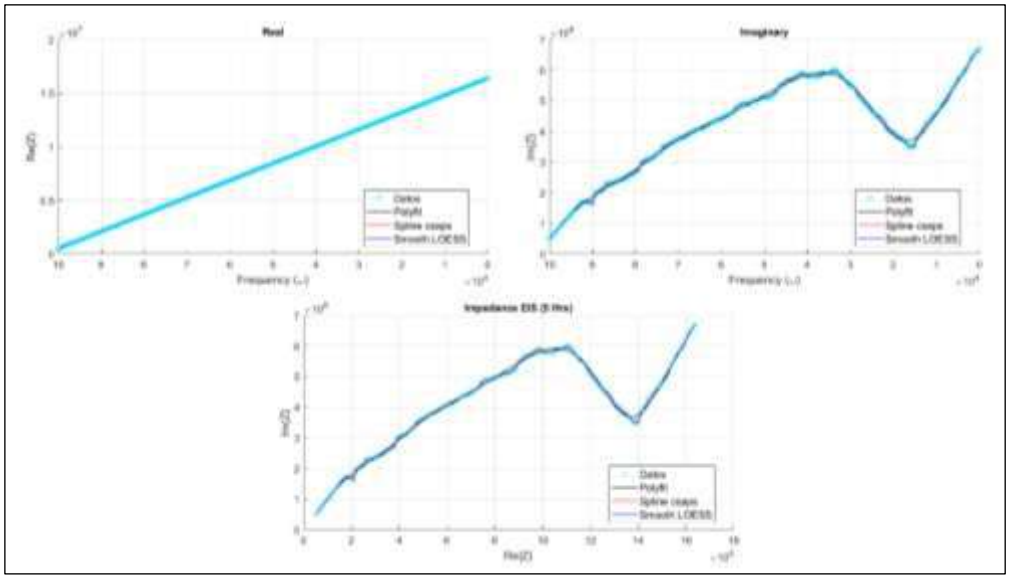


Figure 2: Adjustment of the three methods on real and imaginary impedance data for t = 5 hours.
Source: Authors, (2026).

The real part of the impedance has an almost linear trend with frequency, which means it can be adjusted. Visually, the three methods overlap almost perfectly on the data, which is consistent with the high R^2 values of all models. The imaginary component has a more complex shape with a curvilinear variation and possibly multiple time constants. Both CSAPS and LOESS follow the fluctuations in the data with great fidelity. Polyfit also fits well, but LOESS better captures the changes in curvature observed in the mid frequencies. The Nyquist diagram is the representation most sensitive to the adjustment of the equivalent circuit parameters. The diagram for 5 hours of exposure shows a depressed semicircle, indicating the presence of a constant phase element due to relaxation dispersion in the coating/interface. The overlapping curves show that CSAPS and LOESS offer the smoothest predictions and best fit the pattern of the depressed semicircle, especially at the vertex and in the low-frequency region. Graphically, the large differences in RMSE seen in the tabular results are not apparent, underscoring the importance of numerical metrics (RMSE) for discriminating between visually similar fits.

IV.2 ANALYSIS AND DISCUSSION OF RESULTS FOR PREDICTING IMPEDANCE FUNCTION FOR 24 HOURS

This section analyzes the performance of the three fitting methods (Polyfit, CSPAS, and LOESS) for predicting the impedance function Z(w) of the coating during 24 hours of exposure to the aggressive agent in the system. Table 5 shows the optimal performance achieved by each model when selecting its best parameter (degree, p, or span) in the training (validation) dataset.

Table 5: Parameters and metrics (best) for the three adjustment methods applied for t = 24 hours.

Method	Optimal Parameter	RMSE	R^2
Polyfit	Degree 14	149070.7	0.9807
CSAPS	p = 0.01	177630.8	0.9408
LOESS	span = 0.05	97418.67	0.9905

Source: Authors, (2026).

As at 5 hours, LOESS with a span of 0.05, very local smoothing, achieves the lowest RMSE and the highest R^2. This confirms that LOESS is the method capable of achieving the most accurate fit to the data in its optimal configuration. By prioritizing smoothness (p = 0.001), CSAPS obtains a significantly lower R^2 (0.9408) and a higher RMSE than the other two methods at this stage. This suggests that at 24 hours, the impedance function presents more noise or fluctuations than CSAPS, due to its extreme penalty on curvature, is unable to capture efficiently. Polyfit with grade 14 shows a good R^2 (0.9807), outperforming CSAPS in direct fit. The 10-fold cross-validation evaluation is essential for determining how well the model generalizes to unseen data and how consistent its performance is (low standard deviation), see Table 6.

Table 6: Robustness and generalization capacity of the model for the three fitting methods applied for t = 24 hours.

Method	$RMSE_{mean}$	$RMSE_{SD}$	R^2_{mean}	R^2_{SD}
Polyfit	377985.63	286554.5	0.8124	0.3038
CSAPS	177638.49	3.4020	0.9408	≈0.0000
LOESS	170240.04	88776.41	0.9701	0.0295

Source: Authors, (2026).

The CSAPS model is once again the most robust and consistent with the 24-hour dataset. Its R^2 is 0.0000 and its $RMSE_{SD}$ is only 3.4020. This indicates that although its average accuracy is not the highest (R^2_{mean} = 0.9408), its performance is consistent across different random partitions of the data, which is a critical feature against overfitting.

LOESS achieves the best average R^2 and the lowest average RMSE in cross-validation. This suggests that, on average, it is the method that best generalizes the shape of the impedance, although its high $RMSE_{SD}$ makes it less consistent than CSAPS. Polyfit, with its high degree, 14, exhibits overfitting behavior than at 5 hours. The average RMSE and high standard deviation reveal that the fit is extremely sensitive to data selection, resulting in the worst generalization performance R^2 mean. Table 7 shows the results for the entire dataset (test).

Table 7: Metrics for the three adjustment methods applied for t = 24 hours.

Method	RMSE	R^2
Polyfit	134820.0	0.9876
CSAPS	5.7802	1.0000
LOESS	47469.40	0.9985

Source: Authors, (2026).

The consistency of CSAPS observed in cross-validation is confirmed by an almost perfect fit in the complete dataset of RMSE and $R^2 \approx 1.00$. Although the curvature penalty mechanism initially limited its R^2 in the benchmark, it proves to be the most effective strategy for a precise and robust final fit. Figure 3 shows the fit of the three methods to the impedance data at 24 hours of exposure.

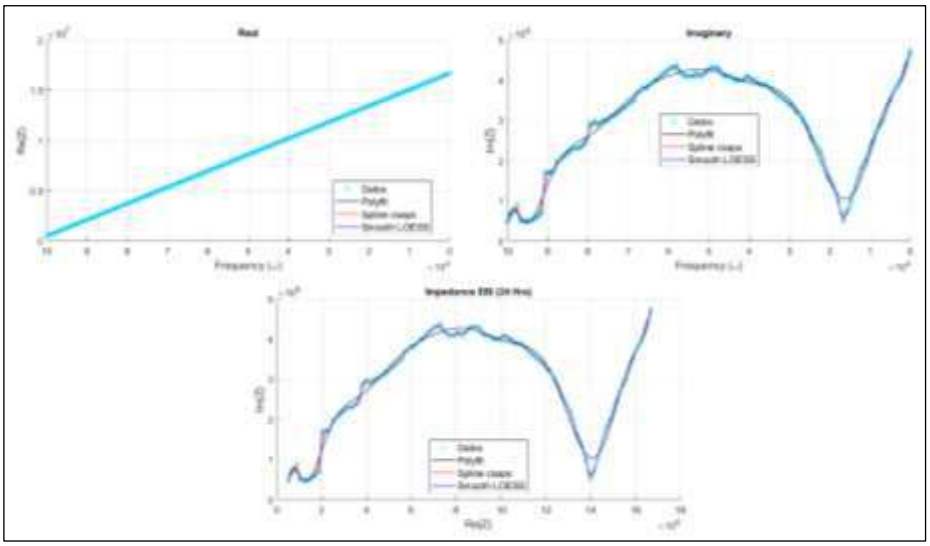


Figure 3: Adjustment of the three methods on real and imaginary impedance data for t = 24 hours.
Source: Authors, (2026).

As at 5 hours, the real part shows a simple trend. The three methods overlap on the data, which is consistent with the high final R^2 values. The imaginary part shows a smoother curve than at 5 hours, with a more defined plateau region between frequencies 7.10^4 Hz and 5.10^4 Hz. All methods follow the trend very well. CSAPS and LOESS fit the slight curvature and detail in the high-frequency flank most accurately. Polyfit follows the trend but may exhibit subtle oscillations inherent in high-degree polynomials. The 24-hour Nyquist diagram shows two distinct arcs, a depressed high-frequency arc and the beginning of a second very low-frequency arc. This is typical of behavior where ion diffusion has already been activated. The CSAPS and LOESS curves trace the pattern of the two arcs with great fidelity and smoothness. The near-perfect fit of the CSAPS suggests that its fitting function passes exactly through the data points, which is visually indistinguishable from the LOESS fit.

V. CONCLUSIONS

The CSAPS method is identified as the most suitable prediction model for the impedance function after 5 hours of exposure, although LOESS achieved the best fit in the benchmark evaluation. CSAPS demonstrated superior robustness and consistency in cross-validation ($RMSE_{SD} \approx 0$), resulting in the most accurate and reliable performance when evaluating the entire data set ($R^2 = 1.00$). The Polyfit method is ruled out as a primary option due to its strong tendency to overfit, evidenced by the high variation in cross-validation performance. For predicting impedance function at 24 hours, the CSAPS method, with its extreme penalty for curvature (p = 0.001), is the predictor of choice due to its excellent performance. Although LOESS demonstrated the highest accuracy in pointwise fitting and the best average generalization ability in cross-validation, the unmatched robustness of ($RMSE_{SD} \approx 3$ and R^2_SD $\approx$ 0.00) led to an almost perfect final fit to the test set ($R^2 \approx 1.00$). This result is fundamental in a scientific context, as it prioritizes reliable and consistent performance across different data subsets. The Polyfit method is excluded because of high variability and the risk of overfitting. Future work will focus on evaluating the proposed modeling strategies across a wider range of EIS datasets, including additional exposure times and coating systems, to assess their robustness under varying degradation mechanisms. Further refinement of the CSAPS and LOESS approaches such as adaptive smoothing parameters or hybrid integration with equivalent-circuit models may improve predictive accuracy and physical interpretability. Additionally, extending the analysis to larger datasets will enable more rigorous statistical validation and strengthen the generalization capability of the models.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Ernesto Bolaños-Rodríguez, Juan Carlos González-Islas.
Methodology: Juan Carlos González-Islas, Asdrúbal López-Chau, Gildardo Godínez-Garrido and Aldo Márquez-Grajales.
Investigation: Ernesto Bolaños-Rodríguez, Asdrúbal López-Chau, Aldo Márquez-Grajales and Fernando Salas Martínez.
Discussion of results: Ernesto Bolaños-Rodríguez, Juan Carlos González-Islas, Asdrúbal López-Chau, Gildardo Godínez-Garrido and Aldo Marquez-Grajales
Writing – Original Draft: Ernesto Bolaños-Rodríguez, Juan Carlos González-Islas and Fernando Salas Martínez
Writing – Review and Editing: Ernesto Bolaños-Rodríguez, Juan Carlos González-Islas, Asdrúbal López-Chau, Gildardo Godínez-Garrido, Aldo Marquez-Grajales and Fernando Salas Martinez
Supervision: Ernesto Bolaños-Rodríguez, Juan Carlos González-Islas.
Approval of the final text: Ernesto Bolaños-Rodríguez, Juan Carlos González-Islas, Asdrúbal López- Chau, Gildardo Godínez-Garrido, Aldo Marquez-Grajales and Fernando Salas Martinez.

VII. ACKNOWLEDGMENTS

We would like to thank our colleagues at the Corrosion Laboratory of the Chemical Engineering Department at the School of Industrial and Mining Engineering at the University of Vigo in Galicia, Spain, for their valuable collaboration in conducting the experimental tests using the electrochemical impedance spectroscopy technique.

VIII. REFERENCES

[1] E. Bolaños-Rodríguez, S. Gil-Fundora, and W. Francisco-Martín, "Determination of the protection behavior of coatings on mirrors using the Electrochemical Impedance Spectroscopy technique," *Journal of Metallurgy*, vol. 39, no. 3, pp. 210–214, 2003, doi: 10.3989/revmetlm.2003.v39.13.331.

[2] I. Obot and I. Onyeachu, "Electrochemical frequency modulation (EFM) technique: Theory and recent practical applications in corrosion research," *Journal of Molecular Liquids*, vol. 249, pp. 83–96, 2018, doi: 10.1016/molliq.2017.11.006.

[3] B. Ramírez Barat and E. Cano Díaz, "In situ evaluation of protective coatings for metallic cultural heritage using electrochemical impedance spectroscopy," *Ge-Conservation*, vol. 8, pp. 6–13, 2015, doi: 10.37558/gec.v8110.278.

[4] S. Papavinasam, R. Rebok, L. Yeng, and N. Berki, *Advances in Electrochemical Techniques for Corrosion Monitoring and Laboratory Corrosion Measurements*. West Conshohocken, PA, USA: ASTM International, 2019, doi: 10.1050/STP1506

[5] A. Cristoforetti, S. Rossi, F. Deflorian, and M. Fedel, "On the Limits of the EIS Low-Frequency Impedance Modulus as a Tool to Describe the Protection Properties of Organic Coatings Exposed to Accelerated Aging Tests," Coatings, vol. 13, no. 3, pp. 598, Mar. 2023, doi: 10.3390/coatings13030598.

[6] X. Liu, W. Li, and Y. Song, "Electrochemical Impedance Spectroscopy Analysis of Coating Degradation Using Advanced Data-Smoothing and Curve-Fitting Approaches," Electrochim. Acta, vol. 389, pp. 138–149, Jan. 2021, doi: 10.1016/j.electacta.2021.138149.

[7] B. Py, "Gaussian-processes for the analysis of electrochemical-impedance spectroscopy data: prediction, filtering, and active learning," ChemRxiv, 2022, doi: 10.26434/chemrxiv-2022-rs2m8

[8] S. Zhao, Y. Jiang, W.-P. He, et al., "Detrended fluctuation analysis based on best-fit polynomial," Frontiers in Environmental Science, vol. 10, art. 1054689, Nov. 2022, doi:10.3389/fenvs.2022.1054689

[9] A. Araveeporn, "The Estimating Parameter and Number of Knots for Nonparametric Regression Methods in Modelling Time Series Data," Modelling, vol. 5, no. 4, pp. 1413–1434, 2024, doi: 10.3390/modelling5040073

[10] S. O. Adams and O. J. Asemota, "The Efficiency of the Proposed Smoothing Method over the Classical Cubic Smoothing Spline Regression Model with Autocorrelated Residual," Journal of Mathematics and Statistics Studies, vol. 4, no. 1, pp. 26–40, 2023, doi: 0.31586/jml.2023.618

[11] M.-H. Chang, Y.-C. Huang, Y.-H. Cheng, C.-T. Terng, J. Chen y J. C. Jan, "Revisiting regression methods for estimating long-term trends in sea surface temperature," Nat. Hazards Earth Syst. Sci., vol. 24, pp. 2481–2494, 2024, doi:10.5194/nhess-2023-218

[12] L. H. Kushi, A. Zimbalist, K. H. Radimer, I. J. Ergas, J. M. Roh, C. P. Quesenberry y M. L. Kwan, "Utilization of locally estimated scatterplot smoothing (LOESS) regression to estimate missing weights in a longitudinal cohort of breast cancer patients," Ann. Epidemiol., vol. 104, pp. 55–60, 2025, doi: 10.1016/j.annepidem.2025.02.013

[13] C. Pérez, *Statistical Sampling Techniques*. Madrid, Spain: Garceta Grupo Empresarial, 2019.

[14] T. M. Dutschmann, L. Kinzel, A. ter Laak et al., "Large-scale evaluation of k-fold cross-validation ensembles for uncertainty estimation," J. Cheminformatics, vol. 15, art. 49, 2023, doi: 10.1186/s13321-023-00709-9

[15] D. Chicco and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Comput. Sci., vol. 7, e623, Jul. 2021, doi: 10.7717/peerj-cs.623.

[16] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," Geosci. Model Dev., vol. 15, no. 14, pp. 5481–5487, 2022, doi: 10.5194/gmd-15-5481-2022

[17] K. Li, "Informativeness of Performance Measures: Coefficients or R-Squareds?", J. Risk Financial Manag., vol. 17, no. 11, art. 481, Oct. 2024, doi: 10.3390/jrfm17110481

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 896-905. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3125

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

ESTIMATING THE SERVICE LIFE OF CONCRETE REINFORCED WITH ADVANCED MATERIALS: INTEGRATION OF IOT SENSORS AND DIGITAL TWINS

Neyser Pérez Pérez[1], Sócrates Pedro Muñoz [2], Víctor Manuel Valdiviezo Sir[3]

[1, 2, 3]Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas - UNTRM, Amazonas, Perú

[1]https://orcid.org/0009-0005-7623-8554, [2]https://orcid.org/0000-0003-3182-8735, [3]https://orcid.org/0000-0003-4348-3076

Email: 7360730192@untmr.edu.pe, socrates.munoz@untrm.edu.pe, victor.valdiviezo@untrm.edu.pe

ARTICLE INFO

Article History
Received: December 12, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Monitoring the condition of the structure,
Built-in sensors,
Concrete,
Fault identification,
Smart infrastructure.

ABSTRACT

The protection of durability and safety in concrete structures increasingly depends on structural health monitoring (SHM) using embedded sensors, which enable real-time monitoring of the internal conditions of the material and support preventive maintenance strategies. This paper analyzes the behavior and adaptability of different sensors integrated into concrete for early damage detection and the development of more robust smart infrastructures. A systematic review of the literature was carried out in the Scopus and Science Direct databases, examining 71 peer-reviewed scientific articles published between 2022 and 2025 in the fields of engineering and materials science and written in English. The findings show that fiber optic, piezoelectric, capacitive, and resistive sensors are particularly effective in recording deformation, temperature, humidity, pressure, and chloride penetration. Furthermore, their combination with artificial intelligence (AI), the Internet of Things (IoT), and digital modeling tools such as BIM and ANSYS increases predictive capability and improves the representation of structural behavior. Similarly, both laboratory research and field applications demonstrate their reliability, resistance to environmental conditions, and potential for scaling. Overall, it is concluded that sensor networks.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Concrete has established itself as one of the most widely used materials in modern construction thanks to its low cost, high strength, and ability to adapt to multiple geometries and structural sizes [1]. The incorporation of advanced sensors has significantly increased the possibility of monitoring concrete, allowing continuous and accurate tracking of its structural condition [2]. Over time, concrete structures can suffer various types of deterioration, such as corrosion, fatigue, or damage from environmental factors, making constant monitoring essential to ensure their integrity [3]. According to [4], these sensors are capable of recording variables such as deformation, temperature, humidity, and pressure, thus facilitating the early detection of possible structural failures. In the field of structural monitoring of concrete, different types of sensors are used, including fiber optic, piezoelectric, and pressure sensors [5].

These devices provide essential information about deformation and internal stresses that accumulate over time, allowing for early identification of cracks or unusual structural behavior [6]. Thanks to these technologies, it is possible to obtain accurate data on the response of concrete to both static and dynamic loads, which is essential for properly guiding maintenance and repair work [7]. According to [8], the use of advanced sensors in concrete monitoring improves accurate fault detection and reinforces preventive maintenance strategies. Early identification of defects in structures using sensors has established itself as a highly effective strategy, as it significantly reduces maintenance and repair costs by enabling interventions before damage reaches an advanced stage [9]. This technology also promotes sustainability in the construction sector, as it facilitates continuous monitoring of structures, extends their useful life, and reduces the need for major repairs or complete reconstruction [10].

The effectiveness of structural monitoring depends largely on the quality of the data collected, its correct integration with forecasting systems, and the ability to generate alerts sufficiently in advance [11]. In this regard, technological advances have enhanced

information analysis and processing tools, making it possible to detect and predict structural failures even before they become visible [12]. Likewise, the incorporation of artificial intelligence and machine learning algorithms has increased the accuracy of predictions, simultaneously optimizing maintenance management in critical infrastructures [13]. The incorporation of sensors in IoT-based systems enables real-time data collection, allowing for continuous monitoring with a high level of accuracy [14]. This massive collection of information provides a solid basis for identifying patterns in structural behavior and anticipating potential failures [15]. According to [16] The integration of these sensors with remote monitoring platforms has strengthened the control of large-scale infrastructure, significantly reducing the costs associated with inspections.

As infrastructure ages, it is essential to monitor it constantly to detect progressive deterioration that is not visible to the naked eye [17]. Regular inspections using specialized sensors enable damage such as internal concrete corrosion to be identified at an early stage [18]. Advances in detection technologies have made it possible to supplement concrete monitoring with methods such as thermography and infrared imaging, which facilitate the identification of anomalies that, in many cases, are not detected by conventional sensors [19]. The combination of these tools provides a more comprehensive assessment of structural condition and increases accuracy in detecting defects such as hidden cracks or the presence of accumulated moisture, factors that could compromise the stability of the structure [20]. Likewise, recent studies indicate that, in the future, structural monitoring will rely on distributed sensor networks, which will optimize the process of supervising and evaluating the condition of structures [21]. With the evolution of these systems, monitoring is expected to become more efficient and accessible, enabling timely and less costly interventions [22].

II. METHODOLOGY

This review article was developed with an exploratory approach, focusing on analyzing scientific publications on the use of sensors in concrete for structural monitoring. The information was obtained through searches in Scopus and ScienceDirect covering 2022–2025. The retrieval strategy used English keywords related to the topic—in particular, the English equivalent of “sensors applied in concrete for structural monitoring”—combined with the Boolean operator AND to refine the accuracy. A search was conducted in the Scopus database using the keywords “sensors applied in concrete for structural monitoring” together with the AND operator. In the first stage, 5,447 documents were identified. After applying filters to restrict the output to scientific articles in English on Engineering and Materials Science, the total was reduced to 2,327 articles, of which 25 were finally selected for analysis. A search was also conducted in ScienceDirect using the same keywords and criteria, with an initial result of 28,380 documents. Subsequently, similar filters were applied (fields of Engineering and Materials Science, type of scientific articles, English language), reducing the total to 10,800. After removing duplicates and reviewing titles and abstracts, 42 articles were selected for analysis and synthesis. The findings are summarized in Tables 1 and 2.

Table 1: Summary, criteria, and results of the search in the Scopus and Science Direct databases.

DATABASE	SEARCH TERMS WITH OPERATORS	ECONRED DOCUMENTS	PERIOD	FILTERS APPLIED	LEAKED DOCUMENTS	SELECTED DOCUMENTS
SCOPUS	sensors Y applied Y to Y concrete Y for Y structural monitoring	5 447	2021 - 2025	Field: “Engineering” “Materials science” Document type: Article	2 327	25
SCIENCE DIRECT	Sensors embedded in concrete for structural monitoring	28 380	2022 - 2025	Field: “Engineering Document type: Article	10 820	42

Source: Authors, (2026).

Table 2: Distribution of article references by year of publication and database.

DATABASE	YEAR OF APPLICATION				
	2022	2023	2024	2025	Total
SCOPUS	1	4	11	9	25
SCIENCEDIRECT	1	7	16	18	42
TOTAL	2	11	27	27	67

Source: Authors, (2026).

Bibliometric analysis and network visualization

PTo determine the state of the art prior to technical synthesis, a search was conducted in Scopus using the unified query:

(TITLE-ABS-KEY (sensors) AND TITLE-ABS-KEY (applied) AND TITLE-ABS-KEY (in) AND TITLE-ABS-KEY (concrete) AND TITLE-ABS-KEY (for) AND TITLE-ABS-KEY (structural) AND TITLE-ABS-KEY (monitoring))

The search returned 616 documents. The metadata were exported and a co-occurrence map was constructed in VOSviewer. Figure 1 shows the network visualization: node size (frequency), color (cluster), and link thickness (strength of association)).

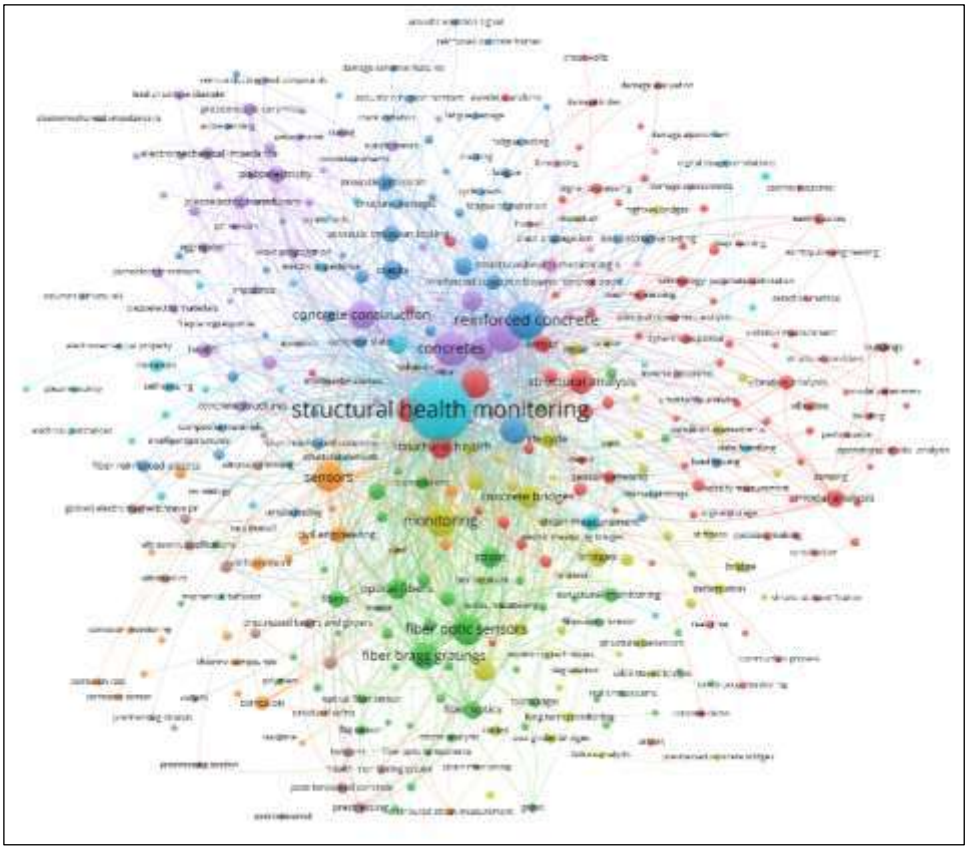


Figure 1: Network visualization.
Source: Authors, (2026).

The graph shows a field structured around structural health monitoring as its core, connected to (i) materials/elements and application scenarios—concrete, reinforced concrete, concrete bridges/buildings, prestressed concrete—; (ii) sensing technologies—fiber optic sensors/DFOS, FBG, piezoelectric/PZT, electromechanical impedance, acoustic emission, digital image correlation—; and (iii) evaluation tasks—damage/crack detection, strain measurement, modal/vibration analysis, condition assessment, corrosion monitoring. Likewise, threads of data management and computation emerge—machine learning, deep learning, computer vision, IoT, and wireless sensor networks — which structure a subfield oriented toward the automation of diagnosis. Overall, the network suggests framing the review around two complementary axes: (a) the typology of sensors and measurement methods and (b) the diagnostic functions on concrete structure typologies, with special emphasis on the early detection of damage, cracks, and corrosion for maintenance and structural safety decisions.

III. RESULTS AND DISCUSSIONS

III.1 INTEGRATED SENSORS FOR STRUCTURAL MONITORING

Integrated sensors allow real-time analysis of the internal response of the material. In the case of self-detectable sensors made of engineered cementitious composite (ECC) and incorporated into columns, they were found to be highly accurate under moderate loads without affecting structural integrity. Through simulations in ABAQUS and measurements taken with strain gauges, it was determined that the most suitable geometry is a 20 mm cube with a friction coefficient equal to or greater than 0.6 [23]. In turn [24], they pointed out that this configuration makes it possible to record parameters such as deformations, temperature, humidity, pore pressure, and the appearance of cracks, which are often difficult to detect through surface measurements. [25].

The importance of continuous monitoring from the system configuration phase to material aging was highlighted, which is essential for identifying faults at an early stage. In his research, a piezoresistive sensor equipped with CBTENG managed to generate 113 V at 4 Hz and charge a 10 μF capacitor to 0.32 V in 25 seconds. In addition [26] hey used artificial intelligence (AI) to detect cracks and corrosion processes, demonstrating its effectiveness after applying acoustic filtration. These developments overcome the limitations of surface sensors, as they provide more comprehensive monitoring, with high sensitivity and self-powering capabilities, which improves both the safety and durability of concrete. According to [27] FBG sensors combined with deep neural networks and PCA analysis enable damage prediction in complex structures with a high degree of accuracy, overcoming the difficulties associated with manual labeling.

By [28] They evaluated the electrical impedance (EI) of embedded PZT sensors, demonstrating that, in samples larger than 150 mm, the EI reflects only the properties of the concrete, eliminating edge effects. This supports their standardized use in continuous structural health monitoring (SHM). Taken together, these results show that artificial intelligence and integrated sensors are revolutionizing structural monitoring by providing accurate predictions and reliable data without the need for manual intervention. Their standardization is a key step toward safer, self-monitoring structures. All this information is summarized in Table 3, based on a review of two articles published in 2023, two articles from 2024, and two articles from 2025.

Table 3: Sensors embedded in concrete for structural monitoring.

TYPE OF INTEGRATED SENSOR	MONITORED PARAMETERS	TECHNICAL ADVANTAGES	MONITORING PHASE	REFERENCE
generic integrated sensor	Cracks, humidity, temperature	Allows for complete integration within the structure	From concrete pouring to operational phases	[23]
generic integrated sensor	Cracks, humidity, temperature	Facilitates monitoring from the early stages to the aging of the material	Throughout the structural life cycle	[24]
FBG (fiber optic) sensor	Deformation, temperature	High precision and resistance to electromagnetic interference	From setting to aging	[25]
Piezoelectric sensor (PZT)	Vibrations, impacts	High sensitivity to dynamic stresses	Continuous monitoring under cyclic loads	[26]
Wireless IoT sensor	Multiple (deformation, temperature, humidity, etc.)	Remote measurement and low maintenance requirements	The entire service life of the structure	
Integrated pressure sensor	Pore pressure	Localized measurement of internal pressure	Initial stages and underground conditions	[27]
Integrated capacitive sensor	Internal humidity, temperature	Good response to hygrothermal changes	First stages of configuration	
generic integrated sensor	mechanical and environmental variables	Promotes the development of smart and sustainable infrastructure	Throughout the entire structural service life	[28]

Source: Authors, (2026).

III.2 FUNCTIONAL CHARACTERISTICS OF SPECIFIC SENSORS

III.2.1 Sensitivity and Accuracy in Fault Detection

In [29] they analyzed the behavior of FBG sensors subjected to biaxial stress, showing that transverse deformation affects longitudinal readings, which reduces the accuracy of uniaxial calibrations. For its part, [30] They pointed out that, thanks to their high sensitivity to microdeformations (με) and changes in wavelength, this type of sensor is capable of detecting internal microcracks before they become visible externally. In turn [31] they used integrated piezoelectric sensors (EPS) to monitor the evolution of concrete strength from the first hour of setting to 90 days, recording phase transitions with remarkable predictive accuracy. Among the systems evaluated, the combination of CPM + CE + slag achieved the highest strength. However [32] Under cyclic loads, EPS can reliably identify internal damage and, when integrated with random forest (RF) models, can predict structural strength with 97% accuracy. Taken together, the reviewed studies show that FBG sensors and embedded EPS provide high sensitivity for detecting microcracks and structural modifications at early stages. Their integration with machine learning techniques significantly improves predictive capability, overcoming the limitations of traditional monitoring methods.

III.2.2 Measuring Range and Resolution

By [33] indicated that the performance of sensors embedded in concrete depends directly on their measurement range and resolution, which are fundamental parameters for ensuring reliable structural monitoring. In the case of fiber Bragg grating (FBG) sensors, resolutions of 1 με and 0.1 °C were reported. According to [34] they added that embedded humidity sensors can achieve an accuracy of ±1% RH, which is crucial for the early detection of corrosion processe. According to [35] Piezoelectric and capacitive sensors perform adequately under dynamic loads, achieving resolutions of 0.01 mm and response times of less than 0.1 s. Finally [36] they developed a method for evaluating concrete permeability using built-in resistivity sensors, calibrated using Van Genuchten models and saturation profiles. The results obtained—with values between 10^{-11} and 10^{-13} m/s—showed correlations greater than 90% compared to traditional techniques such as gamma densimetry and mass loss measurement. These findings confirm that embedded sensors allow accurate and continuous monitoring of critical parameters, improving early detection of structural failures.

III.2.3 Response to External Conditions

By [37] they demonstrated that FBG sensors embedded in concrete allow both the thermal and mechanical conditions of the material to be monitored, recording temperatures of up to 72 °C and deformations of 300 με, with sensitivities of 1.2 pm/με and 10 pm/°C, facilitating thermal compensation and the identification of microcracks. On the other hand, [38] it was reported that the integrated piezoresistive sensors maintain signal stability even under extreme thermal cycles ranging from 20°C to 80°C, making them a suitable choice for structures subjected to severe weather conditions. [39] three-dimensional spacers equipped with RFID-MEMS sensors were also developed to measure moisture at the steel-concrete interface (SCI), achieving an accuracy of ±2% up to 90% RH at 50°C; however, a decrease in accuracy was observed when humidity exceeded 93% due to condensation.

According to [40] FBG sensors were integrated with ultrasonic guided waves (UGW) to detect corrosion in steel bars, recording attenuation greater than 40% in L mode (0.1) under 5 V, which allowed damage to be identified and located even at considerable distances from the point of corrosion. Taken together, these studies show that embedded sensors can operate effectively under extreme conditions, providing accurate measurements of critical variables such as temperature, humidity, and corrosion. Their incorporation into exposed structures improves early fault detection, while technologies such as RFID-MEMS and UGW expand the capabilities of non-destructive monitoring. All this information is summarized in Table 4, based on a review of three articles published in 2023, three in 2024, and five in 2025.

Table 1: Functional characteristics of specific sensors.

TYPE OF INTEGRATED SENSOR	MONITORED PARAMETERS	TECHNICAL ADVANTAGES	MONITORING STAGE	REFERENCE
FBG (fiber optic) sensor	Biaxial deformation, microcracks	High sensitivity to microdeformations (με) and wavelength variations	Seit der ersten Anwendung	[29] ; [30]
Piezoelectric sensor (EPS)	Resistance, phase transitions	Detects changes from the first hour to 90 days; 97% accuracy.	Anfängliche Phasen bis zur Aushärtung	[31] ; [32]
FBG (fiber optic) sensor	Deformation, temperature	Resolution of 1 με and 0.1 °C	Kontinuierliche Temperaturüberwachung	[33]
Humidity sensor	Relative humidity	Accuracy of ±1% RH	Während Prozessen, die mit Korrosion verbunden sind	[34]
Piezoelectric/capacitive sensor	dynamic deformations	Resolution of 0.01 mm and < 0.1 seconds	Unter dynamischen Belastungen	[35]
Resistivity sensor	Concrete permeability	Values between 10^{-11} and 10^{-13} m/s, correlation > 90%	Trocknungs- und Imbibitionsprozesse	[36]
FBG sensor	Temperature and thermal deformation	Temperatures reach up to 72 degrees Celsius (300 degrees Celsius) and 300 degrees Celsius. Sensitivity 1.2 pm/with	Von der Konfiguration bis zum Betrieb	[37]
Piezoresistive sensor	extreme thermal cycles	Signal stability from –20 °C to 80 °C	Umgebungen mit hohen Temperaturschwankungen	[38]
RFID-MEMS sensor	Moisture at the steel-concrete interface	Genauigkeit ±2 % HR bis 90 % HR bei 50 °C	Warme und fe|uchte Umgebungen	[39]
FBG Sensor + UGW	Corrosion on steel bars	Dämpfung > 40 % im Modus L (0,1)	Wassereinwirkung und induzierte Spannung	[40]

Source: Authors, (2026).

III.3 CLASSIFICATION AND PERFORMANCE OF SENSORS APPLIED IN CONCRETE

III.3.1 Fiber Optic Sensors.

Distributed fiber optic sensing (DFOS) technology has proven to be highly effective for the structural diagnosis of prestressed concrete bridges, as it allows real-time and highly accurate measurement of deformations of up to ±500 με, displacements of ±2 mm, and crack opening, as reported by [41]. This technique was also applied to flat slabs subjected to the sudden loss of a column, making it possible to identify the appearance of cracks from the outset and monitor their evolution with a sensitivity of ±20 με. In a practical case [42] they installed a 15-meter prestressed beam belonging to the IDA-KI project using DFOS sensors embedded in the concrete, reinforcement, and cables, achieving a spatial resolution of 1 cm and detecting structural anomalies early on. For its par [43] They proposed an analytical method for asphalt pavements based on DFOS and the CPSD algorithm, validated using accelerometers, obtaining a resolution of 1 cm and a modal criterion greater than 0.95. Together, fiber optic sensors enable distributed, high-precision monitoring capable of detecting deformations, cracks, and failures with excellent resolution and sensitivity. Their effectiveness has been corroborated on bridges, slabs, and pavements, surpassing traditional methods and enabling more reliable early detection of structural damage.

III.3.2 Piezoelectric Sensors.

By [44] they developed a Structural Health Monitoring (SHM) system for aeronautical composite materials that integrates embedded piezoelectric sensors with machine learning techniques, achieving high-precision identification of simple and combined failures through time-frequency analysis based on the Gram angular field. For its part, [45] they used EPS sensors together with the EMI technique to evaluate the evolution of the strength of concrete made from electronic waste, detecting variations from the first 2 hours to 28 days. In mixtures with a 15% replacement of coarse aggregate, the strength was reduced from 37.8 to 34.7 MPa, equivalent to a decrease of 8.15%Asimismo, [46] se evaluaron sensores PZT y PEC bajo condiciones de carga, observándose que el PZT presentó un rendimiento superior en su segundo pico de resonancia, mientras que el PEC destacó por encima de los 800 kHz. El estudio encontró correlaciones logarítmicas entre la deformación (hasta un 0,15 %) y la señal registrada, así como correlaciones lineales entre el voltaje (hasta 25 MPa) y la conductancia. Finally [47] integrated PZT sensors into 3D-printed concrete structures, detecting the presence of cracks through impedance variations of up to 30%. It was observed that 2 mm coatings offered greater sensitivity, while 8 mm coatings increased load capacity by 12%. Taken together, these studies demonstrate that piezoelectric sensors represent a versatile and highly accurate alternative for non-destructive structural monitoring, with applications ranging from additive manufacturing processes to curing control and behavior under dynamic loads.

III.3.3 Capacitive and Resistive Sensors for Concrete Monitoring

Capacitive and resistive sensors allow fundamental concrete parameters to be recorded, such as internal moisture, the appearance of microcracks, and changes in dielectric properties, by measuring variations in permittivity or electrical resistivity associated with processes such as curing, saturation, or carbonation [48]. For example [49] integrated RFID-MEMS sensors into three-dimensional spacers to monitor moisture at the steel-concrete interface. In test specimens with water-to-cement ratios of 0.40 and 0.55, cured for 3 and 14 days and subjected to conditions of 50 °C, 65% RH, and 21 °C, the accuracy reached ±2% up to 90% relative humidity, decreasing above 93% due to condensation, confirming its usefulness for assessing the risk of corrosion. For this reason [50] They developed a cementation sensor composed of a double electrochemical layer and a graphene film, characterized by its high sensitivity to the presence of chlorides using EIS, as well as its stable response to variations in temperature and humidity, although with sensitivity to pH.

Finally, [51] Using EIS, it was observed that the system exhibited high sensitivity and stability, with a response independent of temperature and humidity but affected by pH, which led to the proposal of a calibration equation to improve chloride diffusion models. Together, capacitive and resistive sensors provide accurate monitoring of moisture and corrosion processes from the first 48 hours, even under extreme conditions between 0 and 40 °C and saturation levels of 60 to 100%. Their sensitivity to chlorides and compatibility with wireless technologies make them an ideal choice for exposed structures and continuous monitoring without compromising durability.

All this information is summarized in Tables 5, 6, and 7, compiled from a review of one article published in 2022, two in 2023, four in 2024, and four in 2025. Capacitive and resistive sensors enable accurate monitoring of moisture and corrosion from the first 48 hours, even under extreme conditions of 0–40°C and saturation levels of 60–100%. Their sensitivity to chlorides and compatibility with wireless systems make them ideal for exposed structures and continuous monitoring without compromising durability. °C and 60–100% saturation. Their sensitivity to chlorides and compatibility with wireless systems make them ideal for exposed structures and continuous monitoring without compromising durability. All this information is summarized in Tables 5, 6, and 7, based on a review of one article published in 2022, two in 2023, four in 2024, and four in 2025.

Table 2: Fiber optic sensors.

SENSOR TYPE	ACCURACY	COST	LIFESPAN	ENVIRONMENTAL RESISTANCE	RECOMMENDED APPLICATION	AUTHOR(S)
FBG	±500 con ±2 mm	High	Very tall	Excellent	Post-tensioned bridges	[41]
DFOS	Resolución de 1 cm	High	High	Excellent	Prestressed beams	[42]
DFOS + CPSD	1 cm, >0,95 criterio modal	High	High	Excellent	Asphalt pavements	[43]

Source: Authors, (2026).

Table 3: Piezoelectric sensors.

SENSOR TYPE	ACCURACY	COST	LIFESPAN	ENVIRONMENTAL RESISTANCE	RECOMMENDED APPLICATION	AUTHOR(S)
PZT-5H	±10%, up to 500 kHz	Medium	High	Good	crack detection	[44]
EPS + EMI	2h–28d, -8.15% resistance	Medium	Medium	Moderate	Difficult curing and development	[45]
PZT/PEC	0.15% deformation, 25 MPa	Medium	Medium	Moderate	mechanical evaluation	[46]
PZT in 3D	30% impedance, +12% load	Medium	High	Good	3D Concrete Printing	[47]

Source: Authors, (2026).

Table 4: Capacitive and resistive sensors.

SENSOR TYPE	ACCURACY	COST	LIFESPAN	ENVIRONMENTAL RESISTANCE	RECOMMENDED APPLICATION	AUTHOR(S)
Built-in resistance	0–40 °C, 60–100% relative humidity	Low	High	Moderate	Initial curing and corrosion detection	[48]
Cement with graphene	High sensitivity through EIS	Medium	Medium	Good	Chloride diffusion assessment	[49]
Wireless resistive sensor	Stable remote measurement	Medium	Medium	Moderate	Monitoring of structures in service	[50]
Capacitive and resistive sensors	Variations in permittivity and electrical resistivity	Medium	Medium	Moderate	Moisture control, crack detection, curing, and carbonation	[51]

Source: Authors, (2026).

III.4 ADVANCED TECHNOLOGICAL INTEGRATION FOR STRUCTURAL MONITORING

III.4.1 Artificial Intelligence and Machine Learning

The application of artificial intelligence in SHM allows complex patterns to be recognized using algorithms such as neural networks and support vector machines (SVM), achieving accuracies of up to 98% in the detection of structural damage in beams, even when limited data is available [52]. For its part [53] They emphasized the role of smart sensors in structural health monitoring (SHM), highlighting their ability to replace conventional methods, optimize maintenance strategies, and reduce operating costs by up to 30%, in addition to reporting 93% accuracy in detecting microcracks using artificial intelligence techniques. In the case of bridges [54] reported a decrease of up to 35% in maintenance costs thanks to the use of these systems, although they noted that challenges remain related to high energy consumption and the cost of IoT sensors. Finally [55] indicated that automatic structural condition classification allows early warnings to be generated in accordance with current regulations, strengthening failure prevention. Overall, artificial intelligence is positioned as a key tool for transforming simple data collection into more efficient structural decisions. Despite the high accuracy achieved, its performance depends on the quality and quantity of the data used in training, which poses challenges in terms of standardization and mass acquisition of information.

III.4.2 Structural Data Modeling and Visualization

The use of digital twins combined with real-time sensors and convolutional networks such as VGG-16 has made it possible to identify damage caused by saline agents with high precision, significantly improving predictive maintenance strategies, according to [56]. For its pairte [57] they also pointed out that integrating these systems with BIM platforms facilitates the visualization and detailed analysis of structural damage and pathologies. In the same vein [58] They highlighted that interoperability between IoT sensors and BIM, reinforced by the extension of the SAREF ontology through IADOM and RDF, optimizes both maintenance management and infrastructure design in smart environments.

Likewise, [59] They reported that models capable of instantly updating data from sensors within three-dimensional platforms such as ANSYS APDL have been evaluated, enabling efficient synchronization between analytical models and structural monitoring systems. This advanced visualization capability enhances the interpretation of technical information and facilitates informed decision-making; however, its implementation still requires a robust digital infrastructure, which may limit its adoption in contexts with low levels of digitization or connectivity. All this information is summarized in Table 8, based on a review of one article published in 2022, four in 2024, and three in 2025.

Table 8: Technological integration in advanced structural monitoring.

APPLIED TECHNOLOGY	MAIN FOCUS	OUTSTANDING STRUCTURAL APPLICATION	REFERENCE
Artificial intelligence in SHM	Neural networks and SVM	Identification of faults in beams with 98% accuracy	[52]
Smart sensors with AI	Microcrack detection	0% reduction in operating costs and 93% accuracy	[53]
IoT systems applied to bridges	Structural monitoring	35% reduction in maintenance; associated energy challenges	[54]
Structural classification algorithms	Based on current regulations	Issuing early warnings in accordance with structural standards	[55]
Digital twins + CNN	VGG-16 Architecture	Accurate detection of damage caused by saline agents and predictive maintenance	[56]
BIM modeling in damage assessment	Visualización tridimensional	Advanced analysis of complex structural pathologies	[57]
SAREF ontology with IADOM and RDF	Semantic interoperability	Optimized management in smart environments integrated with IoT and BIM	[58]
EF models integrated with sensor data	Real-time updates	Structural synchronization in 3D platforms such as ANSYS	[59]

Source: Authors, (2026).

III.5 EXPERIMENTAL VALIDATION AND CASE STUDIES

III.5.1 Laboratory Applications

By, [60] They demonstrated that integrated sensors enable early identification of corrosion processes in concrete through acoustic emissions close to ~30 dB, associated with deformations of 20 με, a decrease in amplitude (from 2.59 to 1.29 mV), and an increase in ultrasonic velocity up to 4.55 km/s. For its part, [61] they evaluated pH sensors embedded for 250 days in mortars subjected to carbonation, observing failures caused by accelerated drying at pH values above 12.5 and proposing improvements using gel indicators and compounds such as thymolphthalein. Likewise [62] demonstrated that integrated sensors capable of measuring humidity, temperature, and conductivity can anticipate conditions conducive to corrosion up to 20 days in advance, and that UHPC reinforced with 1.5–4% steel fibers increases sensitivity to damage through the use of FCR. Finally, [63] they validated vibration sensors to detect loss of stiffness under cyclic loads, as well as a matrix sensor (4×4 Ag/AgCl and 2×2 MnO_2) capable of monitoring chloride concentrations in marine environments even in the event of partial system failures. Together, the embedded sensors enable early detection of corrosion, carbonation, and various structural damages. Their performance is enhanced when optimized configurations and materials—such as steel fibers or conductive matrices—are used, strengthening the structural monitoring capabilities of concrete.

III.5.2 Field applications

The integrated FBG sensors, used on the Streicker Bridge and on medium-small span bridges, enabled high-fidelity measurements of temperature, deformations, and vertical displacements (linearity 0.998, sensitivity 7.87 pm/mm, and error <5%), with installation angles between 15° and 30° recommended. In urban settings [64] reported that the London Underground incorporated piezoelectric sensors to detect critical vibrations. With this, they developed an automated SHM system based on modal identification, neural networks, and wavelets, capable of locating damage without manual intervention [65]. On the other hand [66] they proposed an automated SHM framework with modal identification, neural networks, and wavelet transforms, achieving accurate damage detection. Finally [67] they designed a piezoelectric sensor with integrated PZT-5H to detect cracks in concrete beams, demonstrating high sensitivity and stability when recording amplitude variations associated with damage. Using wavelet analysis, it was possible to estimate the depth and location of the cracks. Together, these field applications demonstrate the reliability and scalability of integrated sensors, although their optimal performance depends on proper installation, correct maintenance, and intelligent systems that efficiently process data.

IV. CONCLUSIONS

The systematic review confirms the effectiveness of sensors embedded in concrete as fundamental tools for real-time structural monitoring, allowing continuous monitoring from setting to extreme operating conditions, with high precision in measuring deformation, humidity, temperature, pressure, and corrosion. Fiber optic, piezoelectric, capacitive, and resistive sensors stand out over traditional methods due to their greater sensitivity, environmental resistance, and adaptability to the different stages of the structural life cycle. The integration of these sensors with artificial intelligence, machine learning, IoT, BIM, digital twins, and specialized software such as ANSYS significantly improves fault prediction, structural representation, and decision support, driving predictive maintenance and reducing operating costs. Both laboratory studies and field tests validate their reliability, scalability, and efficiency. However, their optimal implementation requires proper installation, algorithmic processing, and the support of robust digital infrastructures, which are essential for the development of intelligent, resilient, and sustainable structures.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Neyser Alcides Pérez Pérez, Sócrates Pedro Muñoz Pérez, Victor Manuel Valdiviezo Sir
Methodology: Neyser Alcides Pérez Pérez, Sócrates Pedro Muñoz Pérez, Victor Manuel Valdiviezo Sir
Research: Neyser Alcides Pérez Pérez, Sócrates Pedro Muñoz Pérez, Victor Manuel Valdiviezo Sir
Discussion of the results: Neyser Alcides Pérez Pérez, Sócrates Pedro Muñoz Pérez,
Writing – Original draft: Neyser Alcides Pérez Pérez
Writing – Review and editing: Neyser Alcides Pérez Pérez, Sócrates Pedro Muñoz Pérez
Resources: Sócrates Pedro Muñoz Pérez
Supervision: Sócrates Pedro Muñoz Pérez, Victor Manuel Valdiviezo Sir
Approval of the final text: Neyser Alcides Pérez Pérez, Sócrates Pedro Muñoz Pérez, Victor Manuel Valdiviezo Sir.

VI. ACKNOWLEDGMENTS

We would like to express our sincere gratitude to the Toribio Rodríguez de Mendoza National University, Chachapoyas, Amazonas, for its institutional support. Special thanks to the ABC Institute (ABCI), Manaus, Amazonas, Brazil, for its valuable experience and contributions throughout the review process. We also appreciate the access provided by Scopus and the ScienceDirect databases, which were essential for compiling the relevant literature for this review. Finally, we would like to thank all the authors whose work has been cited and who have contributed to the development of this comprehensive review.

VII. REFERENCES

[1] X. Zhang, L. Long, W. Broere and X. Bao, "Smart sensing of concrete crack using distributed fiber optics sensors: Current advances and perspectives," Case Studies in Construction Materials, vol. 22, p. e04493, 2025.

[2] S. Poorghasem, H. A. Shah, W. Meng and Y. Bao, "Measuring early-age shrinkage of concrete incorporating shrinkage reducing admixtures using distributed fiber optic sensors," Measurement, vol. 246, p. 116763, 2025.

[3] X. Lin, P. Wang, X. Han, P. Wang, S. Wang and C. Yin, "Quantitative detection of steel bar corrosion in concrete by using capacitive sensor," Sensors and Actuators , vol. 388, p. 116518, 2025.

[4] J. Han, P. Chen, J. Cai, Z. Lai, J. Pan, Z. Song and Z. Li, "Preparation of self-sensing engineered cementitious composites (ECC) sensor and its sensing performances in concrete column," Measurement, vol. 245, p. 116611, 2025.

[5] D. Moser, A. M. Candilejo, L. C. Felgueroso and D. Santillán, "Use of fiber-optic sensors to monitor concrete dams: recent breakthroughs and new opportunities," Structures, vol. 67, p. 106968, 2024.

[6] M. H. Yassin, M. H. Farhat, M. Nahas and A. S. Saad, "Investigation of fiber Bragg grating sensor measurability in concrete beams under static load conditions," Heliyon, vol. 10, no. 22, p. 40105, 2024.

[7] Y. Ma, S. Wang, G. Xin, B. Li, H. Fang, J. Lei, X. Du, N. Wang and D. Di, "A state-of-the-art-review of underground concrete sewage pipelines detection technologies," Measurement, vol. 242, p. 116268, 2025.

[8] T. Jena, A. K. Raj, T. J. Saravanan and T. Bansa, "Deep learning neural networks for monitoring early-age concrete strength through a surface-bonded PZT sensor configuration," Construction and Building Materials, vol. 241, p. 115698, 2025.

[9] R. Gomasa, V. Talakokula, S. K. R. Jyosyula and T. Bansal, "A review on health monitoring of concrete structures using embedded piezoelectric sensor," Construction and Building Materials, vol. 405, p. 133179, 2023.

[10] Z. Wang, V. A. González, Q. Mei and G. Lee, "Sensor adoption in the construction industry: Barriers, opportunities, and strategies," Automation in Construction, vol. 170, p. 105937, 2025.

[11] C. Liu and Y. Wei, ""Experimental investigation on damage of concrete beam embedded with sensor using acoustic emission and digital image correlation"," Construction and Building Materials, vol. 423, p. 135887, 2024.

[12] G. Zhou, F. Tang, G. Li and H. N. Li, "Gold-sputtered LPG fiber sensor pulse electroplated with iron-carbon for monitoring passivation and depassivation of carbon steel in concrete pore solutions," Cement and Concrete Composites, vol. 147, p. 105432, 2024.

[13] A. U. Rehman, I. G. Kim and J. H. Kim, "Towards full automation in 3D concrete printing construction: Development of an automated and inline sensor-printer integrated instrument for in-situ assessment of structural build-up and quality of concrete," Developments in the Built Environment, vol. 17, p. 100344, 2024.

[14] F. Ye, X. Ding, T. Ma, Z. Ye, Z. Su and S. Zhang, "Application research of epoxy resin mortar composite encapsulated piezoelectric ultrasonic sensors in monitoring dynamic modulus of asphalt concrete," Measurement, vol. 246, p. 11664, 2025.

[15] Q. u. Zaman, M. K. Alkharisi, M. Alturki and A. Raza, "Predicting the various modal parameters of reinforced concrete beams for different near surface mounting sensor arrangements," Results in Engineering, vol. 21, p. 101860, 2024.

[16] H. Yamada, R. Yamada, M. Tani and M. Nishiyama, "Damage evaluation for reinforced concrete shear walls via motion capture and optical fiber sensors," Structures, vol. 60, p. 105955, 2024.

[17] S. Liu, Y. Li, Y. Zhang, M. Ma, W. Qiu, L. She, H. Zhou, K. Li and K. Sun, "Monitoring the circumferential cracking of inner concrete core in prestressed concrete cylinder pipe during wire-wrapping using distributed optical fiber sensors," Tunnelling and Underground Space Technology, vol. 157, p. 106330, 2025.

[18] T. Bansal, V. Talakokula and T. J. Saravanan, "Identification of structural parameters in a prestressed concrete beam under chloride-induced corrosion using embedded and smart-probe-based piezo sensors," Measurement: Sensors,, vol. 35, p. 101284, 2024.

[19] Y. Gautam and H. Jebelli, "Design of flexible polyimide-based serpentine EMG sensor for AI-enabled fatigue detection in construction," Sensing and Bio-Sensing Research, vol. 35, p. 100713, 2024.

[20] H. Ibrahim, G. Villain, N. Ranaivomanana, S. Lopes, J. P. Balayssac, T. Devie and X. Dérobert, "Design and validation of a multi-electrode embedded capacitive sensor to monitor the electromagnetic properties in concrete structures," Measurement, vol. 236, p. 115057, 2024.

[21] A. Monaco, G. Bertagnoli, L. L. Mendola, M. C. Oddo and A. Pennisi, "Preliminary validation of an innovative stress sensor for the Structural Health Monitoring of masonry buildings," Procedia Structural Integrity, vol. 15, p. 102457, 2024.

[22] M. Haq, A. Khan, T. Naqvi, M. Yusuf, H. Kamyab and S. Chelliapan, "Parametric analysis of RSB sensors for concrete strength monitoring using hybrid EMI and WP techniques: Numerical investigation," Ain Shams Engineering Journal, vol. 15, p. 102457, 2024.

[23] H. Jinsheng, C. Peng, C. Jingming, L. Zhengcong, P. Jinlong, S. Zhigang and L. Ziyang, "Preparation of self-sensing engineered cementitious composites (ECC) sensor and its sensing performances in concrete column," Measurement: Journal of the International Measurement Confederation, vol. 245, p. 116611, 2025.

[24] L. Ferrando, J. Ramon, M. .. A. Ibernón, R. Zamora, J. Enrique, G. Romero and 2. J. M, "Corrosion Assessment in Reinforced Concrete Structures by Means of Embedded Sensors and Multivariate Analysis—Part 2: Implementation," Applied Sciences (Switzerland), vol. 14, no. 19, p. 9002, 2024.

[25] D. Wenkui, P. Shuhua, W. Kejin, H. Yuhan, S. Long, W. Fan and L. Wengui, "Integrated triboelectric self-powering and piezoresistive self-sensing cementitious composites for intelligent civil infrastructure," Nano Energy, vol. 135, p. 110656, 2025.

[26] V. S. Charlotte, P. Lotfollah and V. Els, "Smart aggregates for acoustic emission monitoring of concrete cracking and reinforcement corrosion," Construction and Building Materials, vol. 443, p. 137644, 2024.

[27] W. Dapeng and Z. Wenda, "Damage detection combining principal component analysis and deep convolutional neural network with dynamic response from FBG arrays," Journal of Civil Structural Health Monitoring, vol. 13, no. 1, pp. 101 - 115, 2023.

[28] M. Duddi, A. Kocherla and K. Subramaniam, "Evaluation of boundary and material influences on the dynamic electromechanical impedance response of the embedded PZT sensor in concrete," Sensors and Actuators A: Physical, vol. 361, p. 114575, 2023.

[29] Martínez, J. Martínez, M. González, F. Terroba and M. Frövel, "Study of the transverse strain effect on the Fiber Bragg Grating Sensor (FBGS) response with polyimide coating under experimental biaxial tests," Composite Structures, vol. 326, p. 117621, 2023.

[30] K. Harkirat and S. Sarita, "Comparison of Surface Bonded Piezoelectric Transducers and Concrete Vibrational Sensors in Damage Detection of Reinforced Concrete Beams," Practice Periodical on Structural Design and Construction, vol. 29, p. 04024060, 2024.

[31] G. Ramesh, T. Visalakshi, R. .. J. Sri and B. Tushar, "Strength monitoring and prediction of blended concrete systems from very early to delayed curing age using embedded piezo sensor data: An experimental and machine learning approach," Journal of Building Engineering, vol. 107, p. 112677, 2025.

[32] J. Phongthorn, W. Pongsak, T. Pithiwat, B. Theerachai, R. Rattiyakorn and .. P. Ruamporn, "Development and evaluation of KNLNTS-cement composites for use as piezoelectric sensors in structural health monitoring applications," Results in Engineering, vol. 25, p. 104526, 2025.

[33] P. Shakor, N. Gowripalan and P. Rocker, "An Innovative Technique to Measure Lateral Pressure of Self-Compacting Concrete Using Fiber Bragg Grating Sensor," SDHM Structural Durability and Health Monitoring, vol. 18, no. 4, pp. 395 - 408, 2024.

[34] A. Jayaraman, M. Vasudevan and P. Vijayanand, "A Novel Embeddable TiO2 Potential Sensor for Corrosion Monitoring in Concrete Structures," Iranian Journal of Science and Technology - Transactions of Civil Engineering, vol. 49, no. 2, pp. 1347 - 1357, 2025.

[35] D. Jianyun, L. Yu, J. Qinghui, L. Yingwei and M. Yongli, "Preparation and characterization of embedded cement-based spherical piezoelectric sensors and their application in acoustic emission monitoring," Gongcheng Kexue Xuebao/Chinese Journal of Engineering, vol. 45, no. 9, pp. 1569 - 1582, 2023.

[36] M. A. Eid, N. Reuge, G. Villain, S. Bonnet and S. P. Lopes, "Monitoring of drying and imbibition of concrete using embedded resistivity sensors for the estimation of permeability," Cement and Concrete Composites, vol. 157, p. 105900, 2025.

[37] H. Jin, D. Young, K. Won and T. jung, "Real-time monitoring of concrete curing using fiber Bragg grating sensors: Strain and temperature measurement," Sensors and Actuators A: Physical, vol. 362, p. 114650, 2023.

[38] M. Adresi and F. Pakhirehzan, "Evaluating the performance of Self-Sensing concrete sensors under temperature and moisture variations," Construction and Building Materials, vol. 404, p. 132923, 2023.

[39] J. Xiang and H. Wong, "Developing a novel non-destructive method to monitor relative humidity at the steel-concrete interface using RFID-MEMS sensors," Results in Engineering, vol. 404, p. 105286, 2025.

[40] N. Habbaba, S. Mustapha and Y. Lu, "Early detection of corrosion in reinforced concrete using ultrasonic guided wave technique correlated with embedded fiber bragg grating strain sensors," Ultrasonics, vol. 154, p. 107701, 2025.

[41] B. Piątek, T. Howiacki, M. Kulpa and T. Siwowski, "Diagnostics of Post-Tensioned Bridge Girders Using Distributed Fiber Optic Sensors," Procedia Structural Integrity, vol. 64, pp. 1581 - 1588, 2024.

[42] B. Richter, D. Messerer, M. Herbers, K. Speck, J. Laukner, C. Gläser, F. Jesse and S. Marx, "Monitoring of a prestressed bridge girder with integrated distributed fiber optic sensors," Procedia Structural Integrity, vol. 64, pp. 1208 - 1215, 2024.

[43] M. Zeng, D. Hauswirth, U. Morf, H. Zhao and A. Puzrin, "Strain mode measurement of asphalt pavements using distributed fibre optic sensor data," Construction and Building Materials, vol. 450, p. 138730, 2024.

[44] W. Deng, R. K. Langat, E. D. Luycker, A. Cantarel and M. Rakotondrabev, "In-situ piezoelectric sensors for structural health monitoring with machine learning integration," Mechatronic, vol. 106, p. 103297, 2025.

[45] G. Kumar, T. Bansal and D. Sharma, "Monitoring hydration and strength development of E-waste concrete: A passive sensing approach using piezo sensors," leaner Materials, vol. 17, p. 100326, 2025.

[46] H. H. Pan and J. C. Guan, "Stress and strain behavior monitoring of concrete through electromechanical impedance using piezoelectric cement sensor and PZT sensor," Construction and Building Materials, vol. 324, p. 126685, 2022.

[47] Q. T. Kai, V. T. Vy, T. Y. Wei, T. M. Jen, W. T. Neng and L. K. Ho, "Real Time Assessment of Smart Concrete Inspection with Piezoelectric Sensors," Case Studies in Construction Materials, vol. 12, no. 18, p. 20799292, 2023.

[48] M. H. L. S. Wan, T. Oyamoto, T. Kanda and K. H. Tan, "Synergistic impact of temperature and pore saturation on corrosion in carbonated reinforced concrete," Case Studies in Construction Materials, vol. 22, p. e04364, 2025.

[49] J. X. Liew and H. Wong, "Developing a novel non-destructive method to monitor relative humidity at the steel-concrete interface using RFID-MEMS sensors," Results in Engineering, vol. 26, p. 105286, 2025.

[50] B. Huang, J. Wang, G. Piukovics, N. Zabihi, J. Ye, M. Saafi and J. Ye, "Hybrid cement composite-based sensor for in-situ chloride monitoring in concrete structures," Sensors and Actuators B: Chemical, vol. 385, p. 133638, 2023.

[51] A. Dinesh, S. Indhumathi and M. Pichumani, "Self-sensing cement composites for structural health monitoring: From know-how to do-how," Automation in Construction, vol. 160, p. 105304, 2024.

[52] R. Katam, V. D. Pasupuleti and P. Kalapatapu, "Machine learning-driven structural health monitoring: STFT-based feature extraction for damage detection," Structures, vol. 78, p. 109244, 2025.

[53] V. P. d. Oliveira, A. Reis and J. A. Alves, "Assessing the Evolution of Structural Health Monitoring through Smart Sensor Integration," Procedia Structural Integrity, vol. 64, pp. 653-660, 2024.

[54] A. Burrello, G. Zara, L. Benini, D. Brunelli, E. Macii, M. Poncino and D. J. Pagliari, "Traffic Load Estimation from Structural Health Monitoring sensors using supervised learning," Sustainable Computing: Informatics and Systems, vol. 35, p. 100704, 2022.

[55] V. M. D. Mucci, A. Cardellicchio, S. Ruggieri, A. Nettis, V. Renò and G. Uva, "Artificial intelligence in structural health management of existing bridges," Automation in Construction, vol. 167, p. 105719, 2024.

[56] J. A. G. Torres, F. J. D. Mota, E. M. A. Guzmán and G. G. Tinoco, "A digital twin approach based method in civil engineering for classification of salt damage in building evaluation," Mathematics and Computers in Simulation, vol. 233, pp. 433-447, 2025.

[57] C. Gragnaniello, G. Mariniello, T. Pastore and D. Asprone, "BIM-based design and setup of structural health monitoring systems," Automation in Construction, vol. 158, p. 105245, 2024.

[58] E. D. Okonta, F. Rahimian, V. Vukovic and S. Rodriguez, "Semantic interoperability on IoT: Aligning IFC and Smart Application Reference (SAREF) sensor data models," Automation in Construction, vol. 177, p. 106328, 2025.

[59] I. Rudenko and Y. Petryna, "Interaction between BIM and FE models in structural health monitoring," Procedia Structural Integrity, vol. 64, pp. 1216-1223, 2024.

[60] R. Guo, X. Zhang, X. Lin and S. Huang, "Active and passive monitoring of corrosion state of reinforced concrete based on embedded cement-based acoustic emission sensor," Journal of Building Engineering, vol. 89, p. 109276, 2024.

[61] A. S. Alemu, K. Ko, J. Jeong, H. Chung, H. Ramézani and H. K. Kim, "Potential and challenges in long-term monitoring of concrete carbonation using embedded-type optical sensors," Case Studies in Construction Materials, vol. 21, p. e03960, 2024.

[62] H. V. Le, T. N. Pham, V. P. Dang, P. L. Dao, D. L. Nguyen, N. T. Tran and D. J. Kim, "Damage sensing characteristics of smart ultra-high-performance concrete containing electrically conductive fiber and particle fillers under high compressive stress," Sensors and Actuators A: Physical, vol. 383, p. 116242, 2025.

[63] T. Xiao, X. Feng, Z. Qu, X. Lu, J. Zhang, R. Zhu, Q. Zhou, A. He, H. Zhang and S. Chen, "An array sensor for monitoring the chloride concentration profile in concrete," An array sensor for monitoring the chloride concentration profile in concrete, vol. 478, p. 141419, 2025.

[64] S. Li, Z. Liang and P. Guo, "Measurement," A FBG pull-wire vertical displacement sensor for health monitoring of medium-small span bridges, vol. 478, p. 141419, 2025.

[65] H. Hasani, F. Freddi and R. Piazza, "AI-driven automated and integrated structural health monitoring under environmental and operational variations," Automation in Construction, vol. 176, p. 106222, 2025.

[66] Q. Peng, X. Wen, H. Jia, Y. Pan, X. Gu, C. Yin and B. Chen, "A novel sensor-independent convolutional neural network for structural damage detection: Illustrated by a case study on gantry crane," Structures, vol. 71, p. 107971, 2025.

[67] B. Fu, P. Li, Y. Zhang, X. wang, S. Zheng and Y., "Research on damage recognition of concrete beams based on embedded piezoelectric sensors," Case Studies in Construction Materials, , vol. 22, p. e04769, 2025.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 906-917. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3126

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

SYSTEMATIC REVIEW OF THE USE OF FLY ASH IN GEOPOLYMER CONCRETE: IMPACT ON ITS PHYSICAL, MECHANICAL, AND MICROSTRUCTURAL PROPERTIES

Cristian Chuquipul Angulo[1], Socartes Pedro Muñoz Perez[2] and Víctor M. Valdiviezo Sir[3]

[1, 2]Professional School of Civil Engineering, Faculty of Civil Engineering and Architecture, Toribio Rodríguez de Mendoza University, Chachapoyas, Peru.
[3] Faculty of Economics and Administrative Sciences, Toribio Rodríguez de Mendoza University, Chachapoyas, Peru.

[1]http://orcid.org/0009-0009-3465-7303, [2]http://orcid.org/0000-0003-3182-8735, [3]http://orcid.org/0000-0003-4348-3076

Email: *socrates.munoz@untrm.edu.pe, victor.valdiviezo.@untrm.edu.pe, 7108101720@untrm.edu.pe

ARTICLE INFO

Article History
Received: December 12, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Concrete
Geopolymers,
Fly ash,
Microstructure
Alkaline.

ABSTRACT

The environmental impact of Portland cement has driven the adoption of alternative binders such as geopolymer concrete (GC) based on fly ash (FA). This review analyzes 69 studies published between 2021 and 2025 that examine how FA influences the physical, mechanical, and microstructural performance of GC. Workability is strongly affected by the spherical morphology of FA and by the concentration of NaOH, with reported slump values ranging from 62 to 220 mm. Alkaline solutions between 12 M and 18 M accelerate the reaction, reducing initial setting times to 30–150 minutes. Mechanically, mixtures incorporating 50–70% FA and supplementary materials such as slag or microsilica can surpass 60 MPa in compressive strength while also improving tensile and flexural behavior. Microstructural analyses using XRD, SEM, and FTIR confirm the formation of N–A–S–H and C–A–S–H gels that refine and densify the matrix. Overall, FA emerges as a sustainable, high-performance alternative for producing GC, with potential for future construction applications.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

In recent years, environmental protection and sustainable development have become one of the most pressing issues for human society[1], In this regard [2] argues that the rapid growth of infrastructure, urbanization, and industry, and similarly in developing countries, around 95% of construction and demolition waste produced is disposed of in landfills or illegally dumped in unoccupied areas [3], This occurs because concrete is a relatively inexpensive and accessible material [4], On the other hand [5] argue that this condition has driven increased demand for concrete over the last few decades, causing a continuous increase in the need for natural aggregates. Therefore, there is concern about the environmental impact of concrete production in terms of the depletion of natural resources [6]. The environmental impact of the construction industry has become increasingly alarming[7], In particular, the cement industry is characterized by its high energy intensity and the presence of multiple hazardous air pollutants, especially during the production of cement clinker[8],

In fact, it is estimated that cement manufacturing contributes to 5-7% of total CO_2 emissions worldwide into the atmosphere[9], because the production of one ton of Portland cement clinker releases between 800 and 1100 kg of carbon dioxide into the atmosphere[10], It is particularly critical that, according to [11], China, the world's largest cement producer, accounted for approximately half of global cement production in 2023, reaching 2 billion tons. The development of green cementitious materials to replace Portland cement (PC) is crucial for sustainable development [12], people have begun to explore the use of alternative cementitious materials [13], In this regard, geopolymers are cementitious materials characterized by a network structure, obtained through the inorganic polymerization of silicoaluminate compounds, which react under alkaline activators [14], Furthermore [15] argue that geopolymers can be produced without calcination, and various industrial solid wastes can be used as precursors.

Geopolymer concrete (GPC) is a substitute for Portland cement (PC) due to its improved mechanical qualities, lower carbon footprint, and ability to use industrial waste as raw materials[16], It also exceptionally withstands freeze-thaw cycles and high temperatures [17]. According to [18] fly ash (FA) generated from coal and biomass combustion represents one of the main industrial by-products worldwide, with FA from municipal solid waste incinerators [19], entering the context of FA. an X-ray diffraction (XRD) analysis has identified relevant crystalline phases such as quartz (SiO_2), mullite ($Al_6Si_2O_{13}$), hematite (Fe_2O_3), and belite (Ca_2SiO_4), which confer pozzolanic and cementitious properties that make them suitable for applications in construction materials[20], However, there is a growing disparity between the supply and demand for FA, driven by its increasing use in the development of sustainable building materials [21], Therefore, it is necessary to move towards a circular economic model that encourages the comprehensive recovery of this waste by promoting its efficient and sustainable reuse within the construction sector [22].

II. METHODOLOGY

This research was based on a review of 69 scientific articles published in high-impact journals indexed in databases such as Scopus and ScienceDirect. The following specific keywords were used: Geopolymer concrete + fly ash, between 2021 and 2025, as shown in Table 01, with the following results: 4 articles in 2021, 9 articles in 2022, 9 articles in 2023, 23 articles in 2024, and 24 articles in 2025, as shown in Table 02. Likewise, the flow of the elicitation process and the parameters established for the search for publications are represented schematically in Figure 01.

Table 1: Research results with filters.

Database	Year of search	Keywords	Document. without filter	Search filter	Document. with Filter	Selected Documents	Total
Scopus	2021-2025	Geopolymer concrete + fly ash	2227	Engineering+ Materials Science+ Article	1185	43	69
ScienceDirect	2021-2025	Geopolymer concrete + fly ash	7707	Engineering+ Materials Science+ Article	1350	26	

Source: Authors, (2026).

Table 2**:** Distribution of articles by database and year of publication.

Database.	Year Articles Published					Total
	2021	2022	2023	2024	2025	
Scopus	2	9	7	7	18	43
ScienceDirect	2	0	2	16	6	26
Total	4	9	9	23	24	69

Source: Authors, (2026).

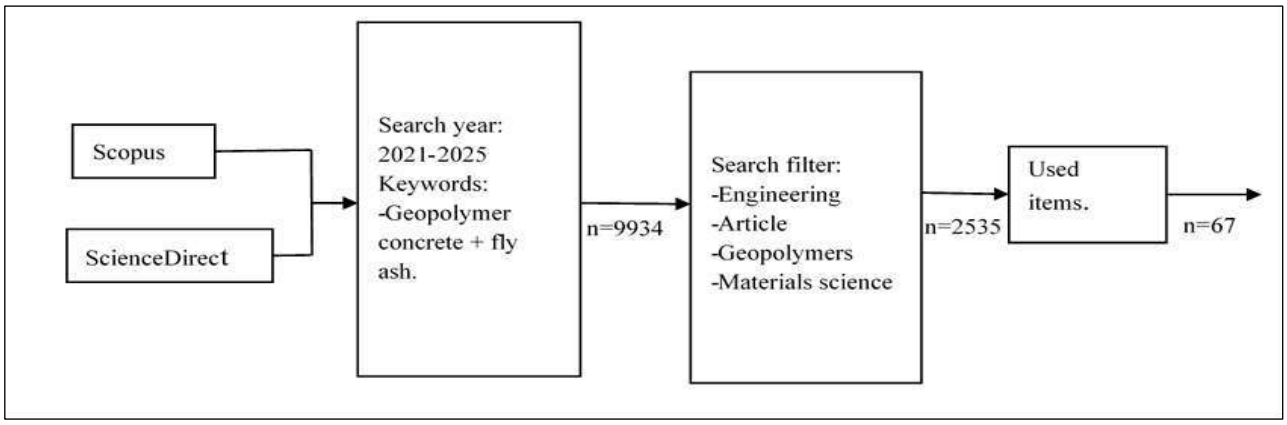


Figure 1: Information Gathering Process Flow and Search Parameters for Publications.
Source: Authors (2026).

II.1 ANALISIS BIBLIOMETRICO

In addition, a bibliometric analysis was performed on the information obtained from the Scopus database, using the following search equation: (TITLE-ABS-KEY (Geopolymer concrete) AND TITLE-ABS-KEY (FLY ASH)) AND (LIMIT-TO (SUBJAREA , "ENGI") OR LIMIT-TO (SUBJAREA , "MATE")) AND (LIMIT-TO (DOCTYPE , "AR")) AND (LIMIT-TO (EXACTKEYWORD , "FLY ASH") OR LIMIT-TO (EXACTKEYWORD , "GEOPOLYMER CONCRETE")) AND (LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2022) OR LIMIT-TO (PUBYEAR , 2023) OR LIMIT-TO (PUBYEAR , 2024) OR LIMIT-TO (PUBYEAR, 2025) OR LIMIT-TO (PUBYEAR, 2026)). This search yielded a total of 1,185 documents, which were analyzed and processed using VOSviewer to generate a keyword co-occurrence map showing the thematic interrelationships within the field of fly

ash use in geopolymer concrete, as shown in Figure 2. The interpretation of the keyword co-occurrence map generated with VOSviewer, based on data extracted from Scopus, allows us to identify the main trends and emerging areas in research on geopolymer concrete.

The results show that the terms “geopolymer concrete” and “fly ash” have greater frequency and centrality, which highlights the dominant role of fly ash as the main precursor due to its abundance and reactivity under alkaline activation. Likewise, lines of study related to sustainability, the use of industrial waste, mechanical and durability properties, as well as the incorporation of artificial intelligence tools for mixture optimization and predictive modeling are identified. Overall, the bibliometric analysis reveals an interdisciplinary evolution linking materials engineering, sustainability, and technology, although gaps remain regarding field-scale application, durability under variable environmental conditions, and the exploration of alternative precursors. Thus, the current trend is toward the development of more efficient, sustainable, and adaptable geopolymer concretes for various construction contexts.

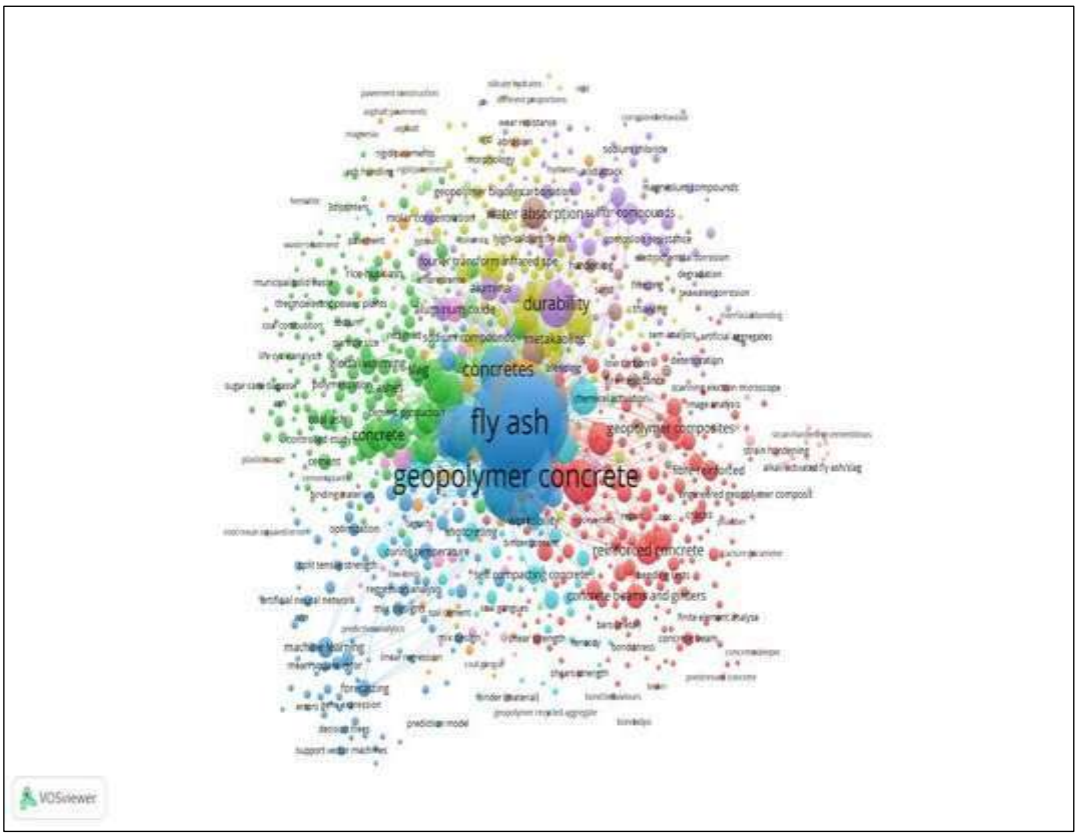


Figure 2: Keyword Co-Occurrence Network Map.
Source: Authors, (2026).

III. RESULTS AND DISCUSSIONS

III.1 PHYSICAL PROPERTIES

III.1.1 Slump

The slump value is tested to determine the workability level of the mixture (GPC), since a higher slump value indicates that the mixture has good workability. In fact, a high slump value is not always accompanied by high compressive strength [23]; In context, the slump of the 0G/100F composition on the (FA) binder is 53.3% lower than that of the 100G/0F composition. By increasing the proportion of (FA) in the combined binder, the slump of the mixtures decreases [24], Likewise [25] highlight that, for (FA)-based RAGPC with a high AAC/B ratio of (0.8) (self-compacting concrete and binder), it was not necessary to use a superplasticizer (SP) to achieve the desired slump. Furthermore, the slump value of the (GPC) increased with increasing molarity of the NaOH solution, with slump values for 6 M, 9 M, 12 M, and 14 M were 62 mm, 78 mm, 87 mm, and 94 mm, respectively [26]. Overall, in mixtures incorporating (FA), workability can be maintained or even improved due to the spherical morphology of its particles, which favors the flow of the paste. Thus, the interaction between the NaOH concentration and the fly ash content defines the balance between the fluidity and cohesion of the material.

In turn [27], There is a slight decrease in the slump of the mixture with sand type M compared to the natural sand mixture, due to the angular shape of the fine aggregates. In another study, the CL1 mixture containing 60% GGBS and 40% pulverized fly ash (PFA) recorded the lowest slump value (185 mm), while the highest value (220 mm) was recorded for the CL3 mixture consisting of 80% GGBS and 20% (PFA). The data indicate that consistency generally improves by keeping the A/P ratio constant and varying the content of granulated blast furnace slag (GGBS) and (PFA) [28], complementing [29] It is argued that the slump test results for GPC with NaOH molarities of GPC12M, GPC14M, and GPC16M are 185 mm, 180 mm, and 175 mm, respectively, indicating that an increase in NaOH molarity leads to a decrease in the workability of fresh GPC. Overall, the studies analyzed show that the workability of GPC is influenced by the physical and chemical factors of its components. All this information is summarized in Table 3, based on a review of one article published in 2022, two in 2024, and four in 2025.

Table 3: Results obtained for the Slump.

Construction material	Activating solution	Percentage of substitution	Settlement (cm)	References
Concrete + (FA)+ Ground coffee	NaOH + Na_2SiO_3	90(FA)%+10% ground coffee	128 mm	[23]
Concrete + (FA) + (GGBS)	NaOH	80(GGBS) %20(FA)%	75mm	[24]
Concrete + (FA)+Recycled Aggregates (AR)	NaOH (SH) + Na_2SiO_3 (SS)	Self-Washing and Bonding Aerated Concrete (ACC/B) =0.8	100mm	[25]
Concrete + (FA)	NaOH + Na_2SiO_3	(CPC)-14M	94mm	[26]
Concrete +(FA) class F + (GGBS)	NaOH + Na_2SiO_3	80%(GGBS)+ 20% (PFA)	220mm	[27]
Concrete + (FA) + (GGBS)	NaOH + Na_2SiO_3	50%(GGBS)+ 50% (FA)	650mm	[28]
Concrete + (FA)	NaOH-12M	(GPC)-12M	185 mm	[29]

Source: Authors, (2026).

III.1.2 Air Content

Geopolymer concrete (GC) was prepared with different silica fume (SF) contents (15%, 22.5%, and 30% by volume). In addition, three Portland cement (PC) concretes with a composition and mechanical properties suitable for comparative purposes were prepared, where the air content of (GC) was lower than that of (PC) and decreased as the (FA) content increased. The air content range of (PC) was maintained between 0 and 0.4% [30], on the other hand [31] It maintains that the minimum air content value observed in the M1R0 concrete mix was 1.50 (±0.5); likewise, the air content for the M1R0, M2R25, M3R50, M4R75, and M5R100 concrete mixes was 1.5%, 1.62%, 1.69%, 1.7%, and 1.74%, respectively. In addition, the angular size of the WMA significantly influences the mixture, as it requires a greater amount of water, which generates more voids; on the other hand, due to the high CaO content in marble, concrete samples are prone to cracking. In summary, the air content in concrete depends on the nature and form of the materials used. The increase in fly ash reduces the occluded air, while marble waste increases it due to its angularity and high calcium oxide content.

III.1.3 Setting Time

By [32] As the silica fume (SF) content increased, the setting times decreased, reflecting the pozzolanic reactivity of SF, which accelerates the geopolymerization process. At a curing temperature of 60°C, the initial and final setting times of the paste (GP) were 53 and 159 minutes, respectively, for an 18M NaOH solution. [33]. On the other hand [34] They used the penetration resistance method and found that the GPC mixtures, with the exception of mixture FA0 containing 0% FA, 60% slag (S), and 40% waste clay brick dust (WCB), reached initial setting between 4 and 5 hours and final setting between 12 and 24 hours. (FA), 60% slag (S), and 40% waste clay brick dust (WCB), achieved initial setting between 4 and 5 hours and final setting between 10 and 10.5 hours; The WCB-based GPC developed in this study showed a high hardening rate and a penetration resistance of 28 MPa in approximately 15 hours or less. In this context, when the A/CE ratio was 0.38 and the (FC) content was 50%, the initial (final) setting times of the geopolymers with alkaline contents of 8%, 10%, and 12% were 30 (115), 90.5 (170), and 204 (295) min, respectively. [35]. This suggests that variations in the A/CE ratio, alkaline activator content, and temperature have a significant impact on the setting and hardening properties of geopolymers. All this information is summarized in Table 4, based on a review of one article published in 2023, two in 2024, and one in 2025.

Table 4: Results obtained for the setting time.

Variable modified	Solución activadora	Activating solution	Initial setting time (min)	Final Setting Time (min)	References
Concrete + (FA)+ (SF)	NaOH + Na_2SiO_3	30%(SF)+ 70%(FA)	150mim	340min	[32]
Concrete + (FA)+ Escoria	Na2SiO3	Relation a/a 0.45	240 min	630 min	[33]
Concrete + (FA) class C+ (FA) class F	NaOH + Na_2SiO_3	50%(CF) A/C=0.38	30mim	115min	[34]
Concrete + (FA)	NaOH + Na_2SiO_3	NaOH-18M Temperature: 60°	53min	159min	[35]

Source: Authors, (2026).

III.2 MECHANICAL PROPERTIES

III.2.1 Compressive Strength

In turn [36] The decrease in compressive strength was more pronounced when only recycled concrete aggregate (RFA) was used instead of natural aggregate (NFA). The rate of increase in strength is significantly higher under hot curing conditions than under room temperature curing conditions. Among the mixtures (GPC) cured at room temperature (25°C), the mixture (8M GC-25) in proportions of (FA) at 95% and slag at 5%, cured at a temperature of 25° and molarity 8M, presented the highest strength at 28 days (63 MPa), while the mixture (2M GC-25) in proportions of (FA) 95% and slag 5%, cured at a temperature of 25° and molarity 2 M, showed the lowest strength (33 MPa). [37], on the other hand [38] It mentions that compressive strength reaches its maximum value of 69.65 at 100 °C, then decreases with increasing temperature.

As the recycled coarse aggregate (RCA) content increases, the residual compressive strength decreases. Similarly [39] argues that partial replacement (FA) with recycled binder generally decreased compressive strength, mortars with up to 60% recycled geopolymer cement (RGPC) exhibited strength values greater than 50 MPa, sufficiently high for structural applications. It has been shown that the use of recycled aggregates and recycled binders tends to reduce the compressive strength of GPC compared to the use of natural materials. However, curing at high temperatures significantly improves mechanical properties, partially offsetting this decrease. By [40], studied the positive impact of class F (FA) on the increase in compressive strength, On the other hand [41] They maintain that the compressive strength at 28 days of the specimens without ultrafine slag (UFS) was 38 MPa, and increased to 65 MPa with a 50% substitution, observing a notable increase in mechanical properties due to the high unit weight and density.

Similarly, the increase in the proportion of geopolymer binder replacement increased the compressive strength values, reaching 65.5 MPa, 68.7 MPa, and 72.5 MPa for the Geo25, Geo50, and Geo75 mixtures, replacing 25%, 50%, and 75% of (OPC) with geopolymer binders, respectively [42]. In addition, the highest compressive strength was achieved by hot air oven curing, with 33.3 MPa, followed closely by accelerated curing (32.1 MPa) and outdoor curing (31.5 MPa). The optimal dosage of (GPC) was determined to be a ratio of 70% (FA) and 30% (GGBS) [43]. Overall, the studies reviewed confirm that increasing the proportion of supplementary cementitious materials, such as fly ash and slag, significantly improves the compressive strength of GPC All this information is summarized in Table 5, based on a review of two articles published in 2024 and six in 2025.

Table 5: Results obtained for compressive strength.

Construction material	Activating solution	Optimal percentage of substitution	Compressive strength at 28 days (MPa)	References
Concrete + (FA)+GGBFS	NaOH + Na_2SiO_3	0% (NCA) + 30% (RFA)	41.4Mpa	[36]
Concrete+ (FA)	NaOH-8M	95% de(FA)	63 MPa	[37]
Concrete + (FA) + Recycled Aggregates (AR)	NaOH + Na_2SiO_3	50% (RCA) y 50%de(NCA)	67.45Mpa	[38]
Concrete + (FA)+ (RGPC)	NaOH + Na_2SiO_3	60%(RGPC)	50 MPa	[39]
Concrete + (FA) + (FA) Bituminous Coal (BCFA)	NaOH + Na_2SiO_3	40%(BCFA) 60%(FA)	38.97MPa	[40]
Concrete +(FA) + (UFS)	NaOH + Na_2SiO_3	50%(UFC) y 50%(FA)	65MPa	[41]
Concrete + (FA) + Steel Slag (SS)	NaOH(L)-12	(SS): 52.5 %, (FA): 22.5 %	75MPa	[42]
Concrete + FA+ (GGBS)	NaOH + Na_2SiO_3	70% de (FA)	33.3MPa	[43]

Source: Authors, (2026).

III.2.2 Tensile Strength

Accordind to [44], They pointed out that GPC consistently exhibits higher tensile strength values, and although it offers a significant advantage in terms of tensile strength during the early stages of curing, this characteristic highlights its superior mechanical performance compared to ordinary Portland cement (OPC). Likewise [45] They mentioned that the proportion of (GGBS) in the mixture reached a maximum of 20%, since above this level, a decrease in tensile strength was observed. On the other hand, the greatest reduction in tensile strength was observed in the mixture with 100% RCA. Adding carbon fiber (CF) increases tensile strength because the load on the first crack does not exceed the fiber's tensile strength. [46].

Finally, the studies revealed that geopolymer concrete (GPCS) manufactured with different percentages of (SF) and (MK) with additional materials showed greater tensile strength in comparison with the reference mixture of (GPFA); In this regard, favorable results were obtained with the incorporation of 20% metakaolin (MK) and 10% SF, which increased the concrete's flexural strength at 28 days by up to 23% and 20%, respectively [47]. The optimal combination of fly ash (FA) with other cementitious materials and complementary reinforcements generates a significant improvement in the tensile strength of geopolymer concrete. All this information is summarized in Table 6, based on a review of one article published in 2021, one in 2022, and two in 2025.

Table 6: Results obtained for tensile strength.

Construction material	Activating solution	Optimal percentage of substitution	Tensile strength at 28 days (MPa)	References
Concrete+ (FA) +(GGBFS)	NaOH + Na_2SiO_3	1:2,5:3,5 (Binder: fine aggregate: coarse aggregate)	4.3 MPa	[44]
Concrete +(FA) +(GGBFS)	NaOH + Na_2SiO_3	80FA%+20GGBFS%	4.35MPa	[45]
Concrete +(FA)+ carbon fibres (FC)+RCA	NaOH-10M y Na_2SiO_3	Precursor concrete + 0.2% de (FC)	3.5MPa	[46]
Concrete + (MK)+(SF)	NaOH + Na_2SiO_3	(GC)+20%(MK) (GC)+10%(SF)	4.83MPa 4.81MPa	[47]

Source: Authors, (2026).

III.2.3 Flexural Strength

The flexural strength evaluation was performed on 28-day specimens, and the results and comparisons with the control mixture (CC) showed that the GPC mixture exhibited a 17.38% increase in flexural strength. [48], In a similar way [49] reported that the highest flexural strength among the modified samples was obtained with the samples containing 60% foamed geopolymer aggregates plus expanded clay (FGA/ECA/P 60), reaching 4,386 MPa, while the lowest was recorded in the FGA 75 sample, with 2,994 MPa. Likewise, the mixture with the best performance in terms of mechanical properties (flexural strength) was composed of 20% microsilica (SF) and 80% fly ash (FA), which significantly improved its mechanical properties [50]. Finally, another study observed an upward trend in flexural strength values as the ordinary Portland cement (OPC) content increased, reaching maximum values of 12.1, 13.12, and 15.21 MPa at 7, 14, and 28 days, respectively, for the 30% OPC and 70% FA mixture; however, a decrease in (FS) was observed after this maximum [51]. In summary, the flexural strength of geopolymer concrete improves with the increase and appropriate combination of supplementary cementitious materials. The presence of fly ash, microsilica, and OPC promotes matrix densification and strength development.

In turn [52] argues that combining recovered fly ash (RFA) with calcium-rich precursor materials such as ground granulated blast furnace slag (GGBFS) to produce geopolymers is a better alternative to geopolymers that incorporate only RFA. In a complementary manner [53] demonstrated by activating the hydration rate and improving the early strength qualities of class C (FA) concrete, an alkaline solution composed of sodium hydroxide and sodium silicate with an optimal molarity of 12M was added. Similarly, the control sample (C-GS) containing fly ash, slag, alkaline activator, coarse and fine aggregates, water, and a superplasticizer achieved the highest flexural strength, 5.1 MPa, while the sample with the highest plastic content (1.5-GS) obtained a reduced strength of 3.7 MPa, representing a decrease of 26.7% [54]. Finally, the results of the experimental flexural capacity of beams reinforced with longitudinal and transverse steel (GPC) are comparable to those of conventional concrete beams (CCB). The difference between the experimental and theoretical flexural capacities was 7.53% and 6.67%, respectively [55] Overall, the studies reviewed demonstrate that the incorporation of calcium-rich precursor materials and the use of alkaline solutions with optimal molarity significantly improve the strength and flexural behavior of (GPC). All this information is summarized in Table 7, based on a review of one article published in 2024 and seven in 2025.

Table 7: Resultados obtenidos para la resistencia a la flexión.

Construction material	Activating solution	Optimal percentage of substitution	Flexural strength at 28 days (MPa)	References
Concrete +(FA) + (GGBS)	NaOH+ Na2SiO3	40% de (FA)	4.8 MPa	[48].
Concrete +(FA) + light aggregates	NaOH+ Na2SiO3	17.90% de (FA)	4.386MPa	[49]
Concrete +(FA) + (SF)	NaOH+ Na2SiO3	80(FA)%+20%(SF)	7.80MPa	[50]
Concrete +(FA)+(OPC)	NaOH+ Na2SiO3	80(FA)%+20(OPC)%	15.21MPa	[51]
Concrete +(FA) + (GGBS)	NaOH+ Na2SiO3	20%(FA)+80%(GGBFS)	9.4MPa	[52]
Concrete +(FA) high calcium content	NaOH+ Na2SiO3	Na_2SiO_3+NaOH/ (FA): 0.37	7.31MPa	[53]
Concrete +(FA) + GGBS+ Recycled Plastic (PR)	Na2SiO3 ·5H2O	Precursor concrete + 0.5% de (PR)	4.7MPa	[54]
Concrete +(FA) +(GGBS)+ longitudinal and transverse reinforcements	NaOH+ Na2SiO3	Precursor concrete +2-8Ø+3-10Ø	19.84MPA	[55]

Source: Authors, (2026).

III.3 MICROSTRUCTURAL PROPERTIES

III.3.1 Scanning Electron Microscopy (SEM)

By [56], [57] show how the characteristics of aggregates influence the microstructure and properties of materials. Scanning electron microscopy (SEM) and X-ray diffraction (XRD) analysis showed that the use of 2% lime content in steam-cured lime-fly ash geopolymer concrete (L-FGPC) improves the geopolymerization process, resulting in a homogeneous, compact, and finer microstructure and produces a dense gel that reduces the proportion of voids. Therefore, it is believed that lime acted as a catalyst in the L-FGPC mixture when 2% lime content was used. [58]; Furthermore, XRD patterns indicate the formation of well-crystallized materials, with higher peaks of quartz (Q) and mullite (M) noted at 2θ values of 27–32° for all samples. In general, these peaks indicated the formation of the main compound (aluminosilicate hydrate (NASH)) produced by the geopolymerization process [59].

On the other hand [60], XRD analysis allows us to evaluate the effect of municipal solid waste (MFA) replacement rates on (GPC), observing small peak clusters between 18° and 36°, associated with amorphous aluminosilicate gels. These clusters progressively decrease as (MFA) replacement increases, reflecting less hydration product formation. This is due to the lower aluminum and silicon content of MFA compared to MK, which reduces the amount of amorphous gels generated in MFA-GPC. Finally, the XRD results of the geopolymer pastes based on fly ash reinforced with GFRP/FA glass fiber at 3, 7, and 28 days revealed that the crystalline phases of the hardened pastes at different curing ages are mainly quartz (SiO_2), calcite ($CaCO_3$), mullite ($Al_6Si_2O_{13}$), larnite, and portlandite ($Ca(OH)_2$).

Furthermore, the progressive decrease in the quartz peak with curing age indicates the dissolution of Si ions in the alkaline medium during geopolymerization [61]. In summary, XRD analyses show that geopolymerization is strongly influenced by the presence and evolution of amorphous and crystalline phases, demonstrating that the formation of NASH, the variation in quartz and mullite peaks, and the decrease in amorphous gels depending on the type and replacement of the precursor determine both the degree of reaction and the structural stability of the geopolymer. All this information is summarized in Table 8, based on a review of four articles published in 2022.

Table 8: Scanning electron microscopy results.

Condition assessed	Morphology/microstructure observed	Implications	References
Geopolymers of (FA) class F+ Cured with lime and steam.	He observed a dense gel, and the ratio of voids and microcracks decreased.	…..	[58]
Geopolymers of (FA) + nanoclay + nanotitanium	Formation of silicate compound, higher density, decrease in ITZ thickness	Improved ZIT and greater mechanical strength	[59]
Geopolymers of (FA) + (MFA)	Replacement of 10% of MFA continues to appear composed of silicate-aluminate gel, appearance of small cracks.	Decreased structural strength and GPC becomes more porous	[60]
Geopolymers of (FA) class F+ fiberglass (GFRP) + steel slag (SS)	Increase in the amount of gels grouped in the matrix, indicating a high degree of geopolymerization.	Increased mechanical strength	[61]

Source: Authors, (2026).

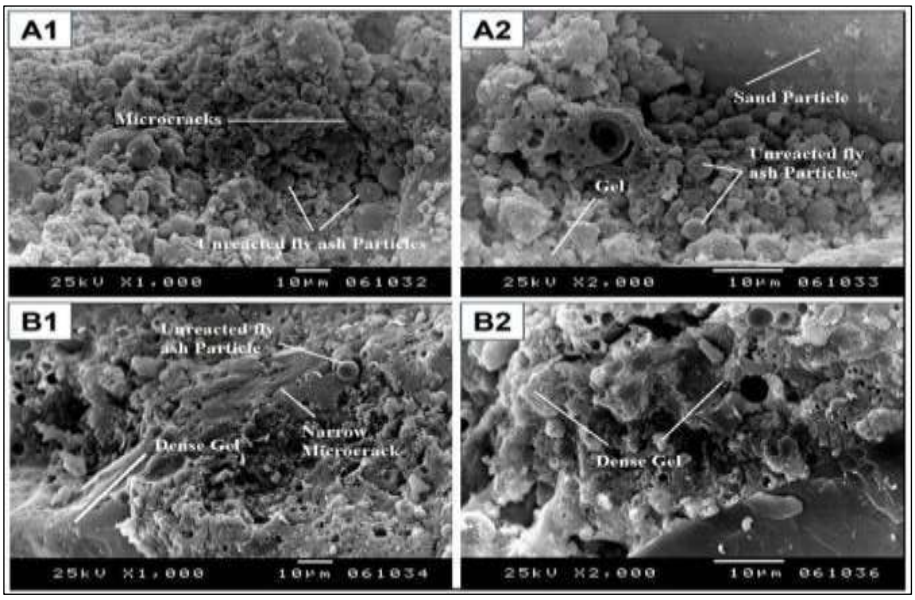


Figure 3: Improvement of the performance of class F geopolymer concrete (FA) through curing with lime and steam.
Source: [60].

III.3.2 X-Ray Diffraction (XRD)

CaO promotes the formation of C(N)–A–S–H gel products, forming a dense microstructure and increasing the compressive strength of self-compacting geopolymer concrete (SCGC) [62]In another study, mixtures A0*, A4, and A8 containing 0%, 4%, and 8% OPC of the total binder at 3 and 28 days, in the early stage, it was found that activation reaction products formed on the surface of the FA particles. This is generally considered to be the precipitation of the sodium aluminosilicate gel produced by the reaction. In contrast, greater precipitation than that mentioned above was observed in sample A8 with 8% OPC [63]. Likewise [64] reported that samples of (GC) 15M2.0NA (natural coarse aggregate, with an NS/NH ratio of 2 and 15 mol) and 15M2.0CM (crushed ceramic waste as coarse aggregate, with an NS/NH ratio of 2 and 15 mol) showed a large crack width and larger pore size in the interfacial transition zone (ITZ) of 15M2.0NA compared to 15M2.0CM.

These results confirmed that the use of waste ceramic aggregate (CM) improved the bonding and homogeneity between the geopolymer paste and the aggregates. Finally [65] They demonstrated that, when comparing the (ITZ) of RS-GP (natural sand and binder) and RFA-GP (recycled fine aggregate and binder), recycled coarse aggregate (RFA) can result in a more compact ITZ than natural sand, which is one of the reasons why compressive strength does not decrease significantly even when the replacement rate of RFA reaches 100%. Overall, studies show that both the presence of CaO and (OPC) and the appropriate selection of aggregates have a decisive influence on gel density, (ITZ) compactness, and, therefore, (GPC) strength. All this information is summarized in Table 9, based on a review of four articles published in 2023.

Table 9: Results obtained for X-ray diffraction.

Condition evaluated	Morphology/microstructure observed	Implications	References
Geopolymers of (FA) + blast furnace slag + (MK)	Reduction in porosity, abundant formation of C–A–S–H gels	Dense microstructure and mechanical strength	[62]
Geopolymers of (FA)+(OPC)	Formation of calcium aluminosilicate hydrate (CASH)	Compressive strength and transverse deformation capacity	[63]
Geopolymers of (FA) + natural fine aggregate (NA) and recycled aggregate (CM)	Greater porosity and water absorption	Lower mechanical properties with (CM)	[64]
Geopolymers of (FA) + natural coarse aggregate (RS) + (RFA)	Lower porosity and water absorption	Lower mechanical properties with (RS)	[65]

Source: Authors, (2026).

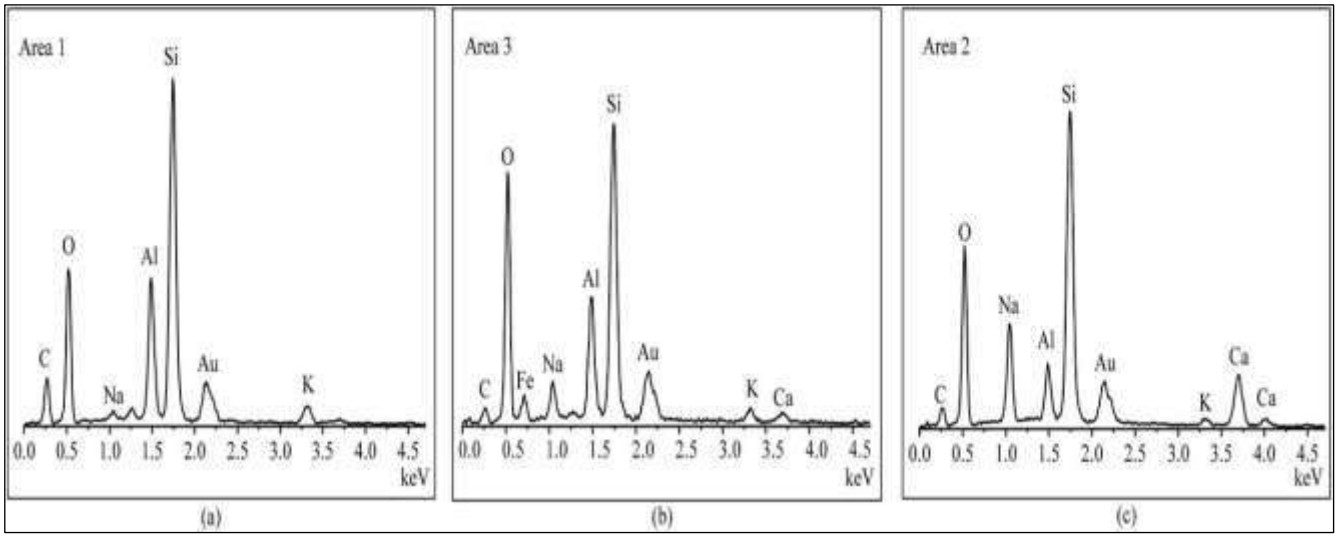


Figure 4: EDS patterns of the FGC matrix at standard curing age. (a): sample A0*; (b): sample A0; (c): sample A8.
Source: [63].

III.3.3 Fourier Transform Infrared Spectroscopy (FTIR)

In GPCs, the optimal CaO/SiO_2 and SiO_2/Al_2O_3 ratios are directly associated with the formation of the C_1S_2H, $C_1(A)_1S_2H$, and $N_1A_1S_2H$ phases, which in turn have a direct influence on microstructure transformation. [66]; according to [67], The FTIR spectra of all specimens consisting of two types of fly ash – high calcium fly ash (HCF) and low calcium fly ash (LCF) – with a weight ratio of HCF to LCF of 1, showed bands at 3700 – 3400 cm^{-1} and 1700–1600 cm^{-1} associated with O–H stretching and H–O–H bending, evidencing the presence of water molecules, while around 1400 cm^{-1} O–C–O stretching was identified in all geopolymers. Likewise, a higher Si–O–Si/Si–O–Al stretching peak was observed in the mixtures with carbon microfiber (CF) compared to the control; however, the mixture with 0.1% (CF) showed a smaller increase in the 1200–950 cm^{-1} band compared to mixtures with higher contents.

Complementarily [68] reported that the FTIR spectra of the CCSW and GCSW groups show similar bands at approximately 3450, 1650, 1450, 1080, 995, and 875 cm^{-1} evidencing a comparable product composition; Furthermore, the absorptions at 1450 and 875 cm^{-1} are mainly associated with carbonates, originating from the calcium carbonate in the concrete slurry (CSW) and a small fraction of carbonized hydrated products. Likewise, the progressive intensification of the CASH band is directly related to the improvement in the mechanical properties of the GCSW samples. Finally, the FTIR spectra of the geopolymers studied in samples Si337, Si325, Si350, and GS15 show that aluminum binds with silicon to form a geopolymer gel, suggesting relatively high compressive strength. in contrast, the peak centered around 972 cm^{-1} in samples GS8, GS4, and Si375 shows that geopolymerization did not generate a solid network and that some slag remained unreacted, especially in GS8 and GS4.

The FTIR results were consistent with the compressive strength of the samples [69]. Together, the FTIR spectra allow the formation and quality of the geopolymer gel to be identified, showing that compositions with a higher degree of polymerization, supported by the intensification of the Si–O–Si/Si–O –Al and CASH bands, develop better microstructures and mechanical properties, while mixtures with lower reactivity show peaks linked to unreacted carbonates or slag, which explains their poor performanceAll this information is summarized in Table 10, based on a review of three articles published in 2021 and one in 2023.

Table 10: Results obtained for Fourier transform infrared microscopy.

Condition evaluated	Identified FTIR bands	Interpretation / Observations	References
Optimal blends: 75%(FA)+25% (UPOFA) 75%(GBFS)+25% (FA)	The shift of (O–Si–O or Si–O–Si) is within 730.95–777 cm^(−1), and the AlO_2 functional group is within 650.12 cm^(−1)–695 cm^(−1).	Formation of gels (C–(A)–S–H) and C–S–H, N–A–S–H	[66]
Geopolymers of (FA) + carbon microfiber (CF)	They showed wave numbers of 3700–3400 cm^(−1) and 1700–1600 cm^(−1) for O–H stretching and H–O–H bending, and O–C–O stretching was detected at a wave number of 1400 cm^(−1).	A noticeable difference was the larger peak elongation of Si–O–Si or Si–O–Al in the CF mixtures compared to the control mixture.	[67]
Geopolymers of (FA) + blast furnace slag (BFS) + concrete slurry (CSW)	The main bands are detected at approximately 3450, 1650, 1450, 1080, 995, and 875 cm^(−1), such as stretching of the –OH group, bending of water (H–OH), carbonate (CO_3^{2-}), silicates or sulfates, and stretching of Si–O.	Gradual improvement of the CASH absorption band leads directly to a significant improvement in the mechanical properties of GCSW samples.	[68]
Geopolymers of (FA)+ (GGBFS)	The bands in the range 900–1200 cm−1 were attributed to Si-OT (T = Si or tetrahedral Al).	A relatively high compressive strength was observed for these samples.	[69]

Source: Authors, (2026).

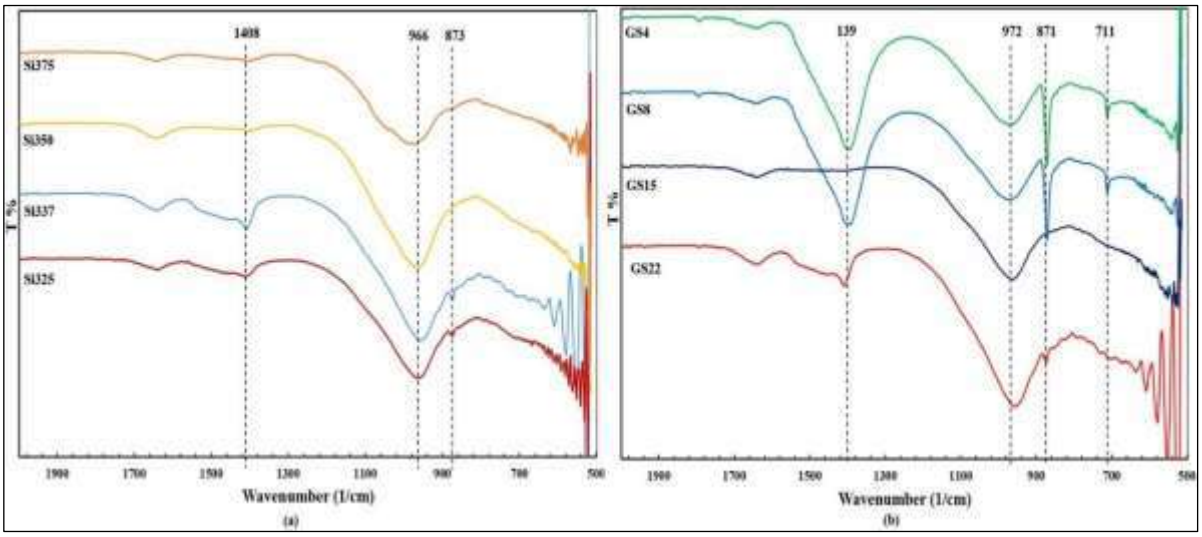


Figure 5: FTIR spectra of geopolymers.
Source: [69].

IV. CONCLUSIONS

Fly ash has established itself as an efficient substitute for Portland cement, reducing the carbon footprint and promoting the circular economy through waste recovery. Its performance in geopolymer concretes depends on proper dosing and the use of alkaline activators such as NaOH and Na_2SiO_3, which generate N–A–S–H and C–A–S–H gels responsible for mechanical and microstructural properties. Factors such as activator molarity, activator-precursor ratio, aggregate type, and curing influence workability and strength development. With the incorporation of supplementary materials such as slag, microsilica, or metakaolin, these concretes can match or exceed conventional concrete in compression, tensile, and flexural strength. Microstructural analyses show dense and stable matrices, confirming that geopolymer concrete with fly ash is a viable and durable alternative for structural and sustainable applications.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Cristian Chuquipul Angulo, Socrates Pedro Muñoz Perez, Victor M. Valdivieso Sir
Methodology: Cristian Chuquipul Angulo, Socrates Pedro Muñoz Perez
Research: Cristian Chuquipul Angulo, Socrates Pedro Muñoz Perez
Discussion of results: Cristian Chuquipul Angulo, Socrates Pedro Muñoz Perez, Victor M. Valdivieso Sir
Original draft writing: Cristian Chuquipul Angulo
Writing: review and editing: Cristian Chuquipul Angulo, Socrates Pedro Muñoz Perez
Resources: Socrates Pedro Muñoz Perez
Supervision: Socrates Pedro Muñoz Perez, Victor M. Valdivieso Sir
Approval of final text: Cristian Chuquipul Angulo, Socrates Pedro Muñoz Perez, Victor M. Valdivieso Sir

VI. REFERENCES

[1] F. Wang, J. Zhai, E. Kan, B. Norkulov and Y. Ding, "Value-added recycling of waste brick powder and waste sand to," Case Studies in Construction Materials, vol. 21, p. e03590, 2024.

[2] S. Filali and A. Nasser, "Evaluating the Impact of Marble Waste and Fly Ash as Sand Replacements on Concrete's Compressive Strength and Workability," Engineering, Technology and Applied Science Research, vol. 14, no. 5, pp. 16797 - 16801, 2024.

[3] A. Younes, E. Elbeltagi, A. Diab, J. Tarsi , F. Saeed and C. Sangiorgi , "Incorporating coarse and fine recycled aggregates into concrete mixes: mechanical characterization and environmental impact," Journal of Material Cycles and Waste Management, vol. 26, pp. 645-668, 2023.

[4] L. Di Sarno, D. Albuhairi and M. J. Peres, "Exploring innovative resilient and sustainable bio-materials for structural applications: Hemp-fibre concrete," Structures, vol. 68, no. 107096, 2024.

[5] H. El-Hassan, P. Kianmehr, D. Tavakoli, A. El-Mi and R.-S. DehKordi, "Synergic effect of recycled aggregates, waste glass, and slag on the properties of pervious concrete," Developments in the Built Environment, vol. 15, p. 100189, 2023.

[6] S. Dias, J. Almeida, A. Tadeu and J. de Brito, "Alternative concrete aggregates - Review of physical and mechanical," Cement and Concrete Composites, vol. 152, no. 105663, 2024.

[7] A. Yonis, Y. Oinam, P. Vashistha, A. B. Degefa, G. . B. Belayneh, S. Park and S. Pyo, "Novel activation method of waste concrete powder for sustainable," Cement and Concrete Composites , vol. 151, no. 105600, 2024.

[8] J. Huang, Q. Jiang, X. Wang, S. Dai, B. Ma and H. Tan, "Effect of silica fume particle dispersion and distribution on the performance of cementitious materials: A theoretical analysis of optimal sonication treatment time," Construction and Building Materials, vol. 212, pp. Pages 549-560, 2021.

[9] Y. Qin, X. Pang, K. Tan and T. Bao, "Evaluation of pervious concrete performance with pulverized biochar as cement replacement," Cement and Concrete Composites, vol. 119, no. 104022, 2021.

[10]Á. M. Sanjuán, E. Menéndez and H. Recino, "Carbonation Resistance of Ternary Portland Cements Made with Silica Fume and Limestone," Buildings, vol. 17, no. 2705, p. 1, 2024.

[11] J. Zhang, Y. Zheng, Y. Zhao, K. Cui, P. Shen and . C. S. Poon, "Utilization of granite sludge in the production of low carbon cement composites after coupled mechanical and CO2 activation (CMCA)," Cement and Concrete Composites, vol. 164, no. 106284, p. 1, 2024.

[12] J. Li, X. Zhang, S. Pang, C. Yang, Y. Wu, L. Su, J. Liu and X. Wei, "Study on the rheological properties and compressive strength mechanism of geopolymer cementitious materials for solid waste," Journal of Building Engineering, vol. 98, no. 111072, p. 1, 2024.

[13] C. Ye, J. Xu and G. Lacidogna, "Fracture behavior of 3D printed geopolymer concrete containing waste ceramic," Compuestos de cemento y hormigón, vol. 163, no. 106193, 2025.

[14] J. Yang , Y. Zhou, P. Du and X. Cheng, "Synchronous vacuum hot pressing ultra-high performance geopolymer: Ultrafast polymerization and early strength development," Journal of Building Engineering, vol. 98, no. 111071, 2024.

[15] S. Wu , X. Yang, S. Xu, D. Chen, Z. Zhao, B. Chen and X. Liang, "Development of a novel emulsified asphalt enhanced steel slag-based geopolymer foamed concrete," Construction and Building Materials, vol. 456, no. 139287, 2024.

[16] I. S. Ayeni , N. H. A. Shukor Lim and M. Samad, "Engineering properties of natural fibre-reinforced one-part geopolymer concrete," Construction and Building Materials, vol. 456, no. 139161, 2024.

[17] G. H. Alcan, A. Benli, A. Oz, B. Bayrak, G. Kaplan and C. A. Aydın, "Effective utilization of silica fume and waste colemanite in eco-sustainable prepacked geopolymers," Construction and Building Materials, vol. 457, no. 139438, 2024.

[18] J. Horvatinec, S. Stipi˘cevi´, S. Bure˘s, M. Brtnicky, J. Holatko, J. Kucerik, J. Buczny and G. Ondrase, "Recycling fly ash from coal or biomass combustion for nutrient recovery and enhanced food production in acidic agroecosystems: A multi-level meta-analysis," Renewable and Sustainable Energy Reviews, vol. 221, no. 115925, p. 1, 2025.

[19] K. S. Liao, W. C. Lu, Y.-L. Chen and Y. Shen, "Production of autoclaved aerated concrete by using municipal solid waste incinerator fly ash as an alternative raw material," Case Studies in Construction Materials, vol. 21, no. e03914, p. 1, 2024.

[20] P. Ferreira , P. Natal, I. Oliveira, A. Bezerra , A. Baldo, A. Silva, J. Dias de Tuesta, J. Peres, D. Ferreira and H. Gomes, "A comprehensive mechanical and physico-chemical characterization of fly ash-based geopolymers," Results in Engineering, vol. 28, no. 108150, p. 4, 2025.

[21] d. Aamar and A. S. Torres, "High-performance geopolymer production using reclaimed fly ash: A sustainable approach to landfilled fly ash management?," Waste Management, vol. 204, no. 114899, p. 1, 2025.

[22] M. Barbato, E. Olsson and L. Rigamonti, "Quality degradation in glass recycling: substitutability model proposal," Waste Management, vol. 182, pp. 124-131, 2024.

[23]M. R. Fadhli and A. A. Setiawan, "A Mixture of Coffee Grounds and Fly Ash in Cementless Geopolymer Concrete," ARPN Journal of Engineering and Applied Sciences, vol. 20, no. pag. 761 - 768, p. 764, 2025.

[24] S. Stelmakh, A. N. Beskopylny, E. M. Shcherban, D. Elshaeva, A. Chernilnik, D. Kuimov, A. Evtushenko and S. Oganesyan, "Geopolymer Concrete Physical and Mechanical Properties on a Combined Binder Reinforced with Dispersed Polypropylene Fiber," Polimers, vol. 17, no. 1710, p. 11, 2025.

[25] B. Gopalakrishna and P. Dinakar, "An innovative approach to fly ash-based geopolymer concrete mix design: Utilizing 100 % recycled aggregates," Structures, vol. 66, no. 106819, p. 12, 2024.

[26] S. Rajpu, S. Rajawat and G. Jain, "Effect of curing conditions on the compressive strength of fly ash-based geopolymer concrete," Materials Today: Proceedings, vol. 103, no. Pages 32-38, p. 37, 2024.

[27] M. Sreenivasa, M. Patil and N. Maya, "Comparative study on geo-polymer concrete using natural sand & M-sand," Materials Today: Proceedings, vol. 62, no. Pages 1879-1882, p. 1880, 2022.

[28] J. Oti, A. Blessing O, L. R. Casabuena, . J. M. Kinuthia and S. Sule, "Utilization of a PFA-GGBS-Based Precursor in Geopolymer Concrete Production as a Sustainable Substitute for Conventional Concrete," Materials, vol. 18, no. 1309, p. 6, 2025.

[29] Y. N. Wibowo, B. Piscesa, Y. Tajunnisa and W. Sutrisno, "Investigation on the mechanical behavior of high-calcium fly ash geopolymer concrete using one-part method," Construction and Building Materials, vol. 490, no. 142436, 2025.

[30] N. Nikoloutsopoulos, A. Sotiropoulou, G. Kakali and S. Tsivilis, "Physical and Mechanical Properties of Fly Ash Based Geopolymer Concrete Compared to Conventional Concrete," Buildings, vol. 11, no. 178, p. 6, 2021.

[31] Saloni, Parveen, Y. . Y. Lim, T. M. Pham, Jatin and J. Kumar, "Sustainable alkali activated concrete with fly ash and waste marble aggregates: Strength and Durability studies," Construction and Building Materials, vol. 283, no. 122795, p. 6, 2021.

[32] P. Verma, D. Goldar, S. . K. Singh and M. Daniyal, "Performance Evaluation and Enhancement of Geopolymer Concrete using Silica Fume," SSRG International Journal of Civil Engineering, vol. 11, no. Pag: 78-88, p. 86, 2024.

[33] B. Gopalakrishna and D. Pasla , "Life cycle assessment (LCA) and the influence of alkaline activator content on mechanical and microstructural properties of geopolymer mortar," Journal of Engineering Research, vol. 13, no. páginas 1462-1474, p. 1466, 2025.

[34] J. Migunthanna, P. Rajeev and . J. anjayan , "Waste Clay Brick as a Part Binder for Pavement Grade Geopolymer Concrete," International Journal of Pavement, vol. 17, no. pages 1450–1467, p. 1457, 2024.

[35]. J. Zhenzhen, X. Li, Q. Yu, Q. Yao and P. Hu, "Sulfate resistance of class C/class F fly ash geopolymers," Journal of Materials Research and Technology, vol. 23, no. páginas 1767-1780, p. 1772, 2023.

[36] S. Irum and F. Shabbir , "Performance of fly ash/GGBFS based geopolymer concrete with recycled fine and coarse aggregates at hot and ambient curing," Journal of Building Engineering, vol. 95, no. 110148, pp. 8,9, 2024.

[37] M. Shamsah, R. Kalfat and K. V. Subramaniam, "Impact of low NaOH molarities on mechanical and durability properties of ambient and oven-cured fly ash geopolymer concrete," Journal of Building Engineering, vol. 105, no. 112491, p. 7, 2025.

[38] C. Jihao, D. Jiayu, Z. Qian, J. Weizhun, L. Chunsheng, L. Mingyan, L. Weihua, L. Yanchun and L. Yajun, "High temperature resistance of slag/fly ash-based geopolymer concrete with fully recycled coarse aggregate," Construction and Building Materials, vol. 482, no. 141587, p. 5, 2025.

[39] A. Naghizadeh, L. Tchadjie, S. Ekolu and M. Welman-Purchase, "Circular production of recycled binder from fly ash-based geopolymer concrete," Construction and Building Materials, vol. 415, no. 135098, p. 5, 2024.

[40] H. Islam, D. Law, Y. Patrisia and C. Gunasekara, "Blended brown coal and Class F fly ash based geopolymer," Case Studies in Construction Materials, vol. 23, no. e05036, p. 7, 2025.

[41] R. Ilayarsi, . K. Mukilan and . M. K. Thiagamani, "Mechanical and durability properties of self compacting fly ash based geopolymer concrete incorporating ultra fine slag," Discover Materials, vol. 5, no. 167, p. 10, 2025.

[42] J. Wang, R. Zhang, X. Zhang, G. Jing and P. Aela, "Static flexure and dynamic behavior of fly ash-steel slag-based geopolymer concrete sleeper," Construction and Building Materials, vol. 495, no. 143642, p. 5, 2025.

[43] S. Pooja and V. Venkatesh, "Influence of accelerated curing on the mechanical and microstructural properties of fly ash–GGBS-based geopolymer concrete with partial recycled aggregate replacement," International Journal of Advanced Technology and Engineering Exploration, vol. 12, no. Pag. 1021-1034, p. 1027, 2025.

[44] Adnan and A. Muhammad, "Geopolymer concrete as a sustainable alternative to OPC," Journal of Umm Al-Qura University for Engineering and Architecture, no. 105647, p. 14, 2025.

[45] L. Sannasi and R. Rajarethinam, "Fresh properties and strength characteristics of ambient geopolymer concrete with ground granulated blast furnace slag," Scielo Brazil, vol. 30, no. e20240795, p. 6, 2025.

[46] P. Nuaklong , A. Wongsa , K. Boonserm , C. Ngohpok , P. Jongvivatsakul, V. Sata, S. Piti and C. Prinya, "Enhancement of mechanical properties of fly ash geopolymer containing fine recycled concrete aggregate with micro carbon fiber," Journal of Building Engineering, vol. 41, no. 102403, p. 6, 2021.

[47] S. Al-Qutaif, S. Ethaib and Y. R. Awei, "Evaluating the Impact of Inclusion Metakaolin and Silica Fume on the Green and Mechanical Properties of Low Calcium Fly Ash Concrete," International Information and Engineering Technology Association, vol. 46, no. pp. 323-331, p. 327, 2022.

[48] S. Mani , P. Partheeban and K. Andiyappan , "Mechanical and durability performance of multilayered hemp–sisal fiber-reinforced geopolymer concrete for sustainable construction," Structures, vol. 81, no. 110328, p. 10, 2025.

[49] A. Przybek, P. Romanska, K. Korniejenko, K. Krajniak, M. Hebdowska-Krupa and M. Łach, "Thermal Properties of Geopolymer Concretes with Lightweight Aggregates," Materials, vol. 18, no. 3150, p. 14, 2025.

[50]. Y. Hamed, M. M. Keshta, . M. M. Elshikh, A. Elshami, M. Matthana and O. Youssf, "Performance of Sustainable Geopolymer Concrete Made of Different Alkaline Activators," Infrastructures, vol. 10, no. 41, p. 12, 2025.

[51] P. Zhou, A. Bahrami, B. Gan and Z. Wang, "Synthesis of Sustainable OPC-Blended Geopolymer Concrete: Experimental and Modeling Study," Advances in Civil Engineering, vol. 2024, no. 7854488, 2024.

[52] A. Danish and A. S. Torres, "Ash to asset: Evaluating the efficacy of reclaimed fly ash in geopolymer production," Journal of Cleaner Production, vol. 503, no. 145326, p. 8, 2025.

[53] S. Rajagopal, S. Rajendran, J. Mylan and G. Murugesan, "A study on high calcium flyash blended geopolymer composite with industrial derivatives," Revista Materia, vol. 29, no. e20240646, p. 14, 2025.

[54] M. Jameel , T. Suteerasak, P. Chumpol, A. Puttiwongrak , W. Sae-Long and P. Sukontasukkul, "Properties of one-part geopolymer pedestrian blocks made using 100 % waste materials," Case Studies in Construction Materials, vol. 22, no. e04538, p. 12, 2025.

[55] K. Anil and M. Shambhu Sharan, "An Experimental and Analytical Study of the Flexural Capacity of reinforced Geopolymer Concrete Beams," Engineering, Technology and Applied Science Research, vol. 15, no. 22109-22113, p. 22112, 2025.

[56] H. Zhong, H. Fu, Y. Feng, L. Li, B. Zhang, Z. Chen, Z. Lu and J. Xie, "Compressive Behaviors of High-Strength Geopolymeric Concretes: The Role of Recycled Fine Aggregate," Buildings, vol. 14, no. 1097, p. 4, 2024.

[57] M. Mudgal, P. K. Goyal, P. K. Ghosh, A. Kumar, R. Chouhan and A. Srivastava , "Performance Assessment of BOF steel slag as fine aggregate in the development of fly ash based geopolymeric mortar," Australian Journal of Structural Engineering, vol. 25, no. 4, pp. 405 - 416, 2024.

[58] M. A. Diab, "Enhancing class F fly ash geopolymer concrete performance using lime and steam curing," Journal of Engineering and Applied Science, vol. 69, no. 59, p. 15, 2022.

[59] N. H. Jumaa, I. M. Ali, M. S. Nasr and M. Falah, "Strength and microstructural properties of binary and ternary blends in fly ash-based geopolymer concrete," Case Studies in Construction Materials, vol. 17, no. e01317, p. 8, 2022.

[60] M. Niu, P. Zhang, J. Guo and J. Wang, "Effect of Municipal Solid Waste Incineration Fly Ash on the Mechanical Properties and Microstructure of Geopolymer Concrete," MDPI, vol. 8, no. 6, p. 15, 2022.

[61] J. Wang, C. Zheng, L. Mo, H. GangaRao and R. Liang, "Assessment of recycling use of GFRP powder as replacement of fly ash in geopolymer paste and concrete at ambient and high temperatures," Ceramics International, vol. 48, no. Pages 14076-14090, p. 14085, 2022.

[62] Y. Luo , Q. Zhang, D. Wang, L. Yang, X. Gao, Y. Liu and G. Xue, "Mechanical and microstructural properties of MK-FA-GGBFS-based self-compacting geopolymer concrete composites," Journal of Building Engineering, vol. 77, no. 107452, p. 15, 2023.

[63] C. Yang, J.-J. You, Y.-W. Huang, X.-M. Ji, Q.-Y. Song and Q. F. Liu, "Low-carbon enhancement of fly ash geopolymer concrete: Lateral deformation, microstructure evolution and environmental impact," Journal of Cleaner Production, vol. 422, no. 138610, p. 9, 2023.

[64] S. Naenudon, A. Wongsa, J. Ekprasert, V. Sata and P. Chindaprasirt, "Enhancing the properties of fly ash-based geopolymer concrete using recycled aggregate from waste ceramic electrical insulator," Journal of Building Engineering, vol. 68, no. 106132, p. 14, 2023.

[65] S. Liu, Z. Liu, K. Takasu, H. Koyamada and H. Suyama, "Study on the Effect of Recycled Fine Aggregate Qualities on Fly Ash/GGBS-Based Geopolymer Mortar," MDPI, vol. 16, no. 23, p. 13, 2023.

[66] E. Otman , A. M. J. Megat , A. A. Zainal , M. Nuha and M. Abdalrhman, "The Compressive Strength and Microstructure of Alkali-Activated Mortars Utilizing By-Product-Based Binary-Blended Precursors," MDPI, vol. 4, no. pp. 885–898, p. 894, 2023.

[67] P. Nuaklong, A. Wongsa, K. Boonserm, C. Ngohpok, P. Jongvivatsakul, V. Sata, S. Piti and C. Prinya, "Enhancement of mechanical properties of fly ash geopolymer containing fine recycled concrete aggregate with micro carbon fiber," Journal of Building Engineering, vol. 41, no. 102403, p. 4, 2021.

[68] Y. Gao, K. Duan, S. Xiang and W. Zeng, "Basic Properties of fly Ash/Slag -Concrete Slurry Waste Geopolymer Activated by Sodium Carbonate and Different Silicon Sources," Frontiers in Materials, vol. 751585, no. 751585, p. 8, 2021.

[69] A. Dehghani , F. Aslani and G. N. Panah, "Effects of initial SiO2/Al2O3 molar ratio and slag on fly ash-based ambient cured geopolymer properties," Construction and Building Materials, vol. 293, no. 123527, p. 8, 2021.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 918-929. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3128

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

MONITORING DIGITAL TWIN FRAMEWORK FOR CONTROLLED ENVIRONMENT AGRICULTURE USING DEEP LEARNING AND EDGE COMPUTING

S. Sasikala*[1] and Sita Devi Bharatula[2]

[1, 2]Department of Electronics and Communication Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Thiruvallur, Chennai, India.

[1]http://orcid.org/0009-0006-2094-2507, [2]http://orcid.org/0000-0001-7022-1026

Email:*s_sasikala@ch.students.amrita.edu, b_sitadevi@ch.amrita.edu

ARTICLE INFO

Article History

Received: December 13, 2025

Revised: January 10, 2026

Accepted: January 15, 2026

Published: February 28, 2026

Keywords:

Controlled Environment Agriculture,
monitoring Digital Twin,
Digix-1 PLC,
Deep Recurrent Q learning model,
SP112-GT40-S-CE HMI
edge computing device.

ABSTRACT

The Controlled Environment Agriculture (CEA) system controls the soil and climatic conditions to enhance agricultural productivity and resource efficiency in a sustainable way. This work offers a framework of the Monitoring Digital Twin (mDT) of CEA operations optimization, which combines automation, real-time monitoring, and predictive analytics. A SELEC DIGIX-1 Programmable Logic Controller (PLC) has automated control of environmental parameters, whereas the digital twin continually gathers sensor data, extracts features, and stores them to be used in the information analysis. The main objectives of the proposed system are (i) to predict Crop yield based on Crop Yield Prediction Dataset and (ii) to detect (weed, pesticide and plant disease) based on Plant Village Dataset. A Median Filter with Z-Score Normalization is used to perform data preprocessing to improve the quality of data and eliminate noise. Gray Level Co-occurrence Matrix (GLCM) is used to obtain effective feature extraction, and then, the Deep Recurrent Q Network model is used with a Convolutional Neural Network (CNN) to classify them. The model has a better performance of an accuracy of 0.97, precision of 0.96, recall of 0.93, F1-score of 0.93, and Root Mean Square Error (RMSE) of 0.1673, of superior performance compared to conventional methods. The data is processed and stored in DynamoDB and hence accessed by Python-based Edge computing devices on user request. In addition, the SP112-GT40-S-CE Human Machine Interface (HMI) displays the insights (locally and remotely) like predicting crop yields, identifying the weed, pesticide suggestions, and identifying the disease in the plant. On the whole, this mDT framework improves the CEA ecosystem by integrating digital twin technology and deep learning in order to attain intelligent automation, early anomaly and sustainable agricultural productivity.

Copyright ©2025 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Controlled Environment Agriculture (CEA) is the emerging alternative to meet the increasing demands for sustainable food production by controlling the soil and climatic environments to ensure optimal agricultural output and resource utilization [1]. The rapid evolution of digital twin (DT) technology has made agriculture even more transformative, as intelligent monitoring, simulation, and control are achieved in various farming conditions. Recent developments in cloud-fog-edge infrastructures have enabled the effective implementation of DTs to support smart agriculture, facilitating real-time data collection, processing, and decision-making [2]. These highly focused surveys have also demonstrated the potential of DT frameworks to enhance crop productivity, reduce the cost of inputs, and mitigate their sensitivity to environmental and biological stresses [3]. In addition to open-field crop production, DT has also been applied to urban arable ecosystems and aquaponics, where decision support tools combine environmental monitoring with predictive analytics to inform optimization options [4]. The integration of DT, edge artificial intelligence (AI), and machine learning (ML) has been identified as one of the main facilitators of intelligent automation and adaptive control in agriculture and industry [5].

In particular, IoT, augmented reality, and DT-based real-time greenhouse management have proven to be efficient and sustainable in their operational capacity and sustainability [6]. Hybrid DT models have also been made to represent greenhouse environment, and the underground environment simultaneously to give more accurate simulations to allocate resources [7]. Similarly, robots that run on DT have demonstrated promising experiments in harvesting crops in a specific manner, such as harvesting strawberries in greenhouses [8]. Intelligent fish farm management systems have been applied to DT-based architectures in the context of aquaculture (based on sophisticated decision-making mechanisms, such as the analytic hierarchy process) [9]. On a larger scale, recent developments integrating DT with collaborative edge-cloud computing and resource allocation algorithms have formed the basis of its implementation in larger 6G-enabled Industrial IoT systems [10]. These developments are evidence that, with the integration of deep learning and edge computing, DT technology represents a revolutionary avenue in contemporary agriculture.

I.1 OBJECTIVE

- To create and deploy a Monitoring Digital Twin (mDT) to unite automation, real-time monitoring and predictive analytics to optimize Controlled Environment Agriculture (CEA) processes based on intelligent decision support.
- To combine Digital Twin technology with benchmark agricultural data, i.e. Crop Yield Prediction Dataset to predict productivity and Plant Village Dataset to detect weeds, pesticides and plant leaf diseases.
- To optimize the quality of the data in the Digital Twin environment with a Median Filter of Z-Score Normalization of noise elimination, normalization, and proper data synchronization between the real and virtual layers.
- To obtain meaningful texture and correlation-based characteristics of the twin generated agricultural data through the Gray Level Co-occurrence Matrix (GLCM) to accurately analyze and map out crop health conditions digitally.
- To create a hybrid Digital Twin model based on deep learning to classify, predict, and detect anomalies in CEA systems using a Deep Recurrent Q Network with a Convolutional Neural Network (CNN).

I.2 CONTRIBUTION OF THE WORK

- Monitoring Digital Twin (mDT) is a proposed architecture of Controlled Environment Agriculture (CEA), which can be used to monitor and predict the environment, automate, and improve crop management, as well as sustainable farming.
- The framework combines two benchmark datasets the Crop Yield Prediction Dataset, which is used to estimate productivity, and the Plant Village Dataset, which is used to detect the presence of weeds, pesticides, and plant diseases, into a single Digital Twin setup.
- The twin data pipeline uses a hybrid method of Median filter and Z-score normalization to remove any noise, normalize the inputs and to increase the accuracy of sensor and image data synchronization.
- The method used to extract the spatial and texture features of agricultural images is called the Gray Level Co-occurrence Matrix (GLCM), which enhances the analytical power of the twin in detecting the health of crops in the images.

A Deep Recurrent Q Network with a Convolutional Neural Network (CNN) is used to classify and predict in the digital twin with very high performance (accuracy, precision, recall, F1-score) and allows a smart and automated decision-making in CEA operations.

I.3 ORGANIZATION OF THE PAPER

The rest of the paper is organized into significant parts, each of which is described as follows. Section II lists the research projects on Monitoring Digital Twin Framework for Controlled Environment Agriculture completed by various authors. The suggested method's workflow is defined in Section III, and the Results and performance analysis of Monitoring Digital Twin Framework for Controlled Environment Agriculture are presented in Section IV. The conclusion of the proposed work that will be done in a future scope is included in Section V, along with references.

II. RELATED WORK

The more recent literature shows a consistent change in the concept of Artificial Intelligence (AI)-based Digital Twin (DT) structures and their application to sustainable and smart agriculture, with each article adding its own methodological nuances and revealing unresolved issues in the field. Based on the principles of the circular economy, [11] combines the concept of Artificial Intelligence (AI) and Digital Twin (DT) models to empower the optimization of resources and waste management in Chinese agricultural industries, where DTs reflect the processes of physical farming, whereas AI helps to make forecasts, but the solution is burdened by an excessive dependence on data and is not proved to be scalable to real-time and across regions. Building on the previous subject, [12] offers an extensive review on AI convergence with DT in agriculture and forestry, including approaches like the use of Machine Learning (ML)-based crop sensors and simulation-based DTs, yet the review is, to a large part, more of a conceptual analysis with little analysis of performance at the implementation level. In the educational field, [13] proposes the use of Extended Reality-based Digital Twins (XR-DT) to supplement agricultural training with immersive visualization and simulation but the platform is hampered by the high hardware costs and lack of integration with actual farm data streams.

Dwelling on the topic of precision farming, [14] addresses the concept of Digital Twin Smart Farming Systems (DT-SFS) based on geospatial and sensor-based agricultural systems, but the study concept does not include a unified architecture of the system and the objective quantification of its accuracy. In the case of aquaculture, [15] applies a Digital Twin Architecture (DTA) to intelligent fish farm operation by ranking system components with the help of a Modified Analytic Hierarchy Process (M-AHP), but the selection of the components is subjective and not tested in dynamic environmental uncertainty conditions. In infrastructure, [16] suggests Cloud-Fog-Edge-based Digital Twin Deployment (CFE-DT) to minimize latency and maximize scalability in smart agriculture, yet the complexity of the architecture will raise the maintenance cost and security-related issues. In [17], the integration of reinforcement learning is known as Reinforcement Learning-based Digital Twins (RL-DT) to adaptive control and optimization but the models have high computational cost and slow convergence in sparse-reward agricultural settings.

In [18], the synergy of Artificial Intelligence (AI) and the Internet of Things (IoT) is highlighted through an AI-IoT Smart Agriculture Framework (AIoT-SAF) of real-time monitoring and decision support, but the interoperability issues, as well as energy consumption of IoT devices is not resolved. The system level validation is provided in earlier [19] that creates a Digital Twin of an Urban Hydroponic Farm (DT-UHF) with data-centric modelling to model crop growth and resource flows; however, it is highly accurate but is specific to the domain and hard to extrapolate to open-field agriculture. In [20], interpretability is considered with the help of Interpretable Artificial Intelligence-based Smart Agriculture System (IAI-SAS) which integrates transparent ML models with decision rules, but because of the trade-off between interpretability and predictive performance, the application is only applicable to more complicated situations. Sustainability wise, [21] is a review of Smart Agriculture Digital Twin Applications (SA-DT) and outlines challenges of data integration, standardization, and lifecycle management, but does not go as far as suggesting specific optimization algorithms.

In [22], a paradigm shift to complete digitalization is discussed, and the concept of Next-Generation Digital Twin Paradigm (NG-DTP) of agriculture is discussed, but the lack of experimental validation of the concept undermines its practical significance. An advanced synthesis in [23] classifies Digital Twin Models in Agriculture (DTMA) according to the functionality and maturity, yet points at the lack of synchronization in real-time and autonomy of AI. Expanding the scope, [24] samples Cross-Industry Digital Twin Applications (CI-DT), which can be transferred to agriculture but has no domain-specific performance metrics. Lastly, [25] provides a detailed overview of Digital Twin-based Smart Agriculture Systems (DT-SAS), with specific focus on sensor fusion, AI analytics, and system integration, but it finds that the problem of scalability, cybersecurity, and cost-effectiveness still have not been resolved.

Table 1: Comparison of Digital Twin Applications in Agriculture

References. No	Author & Year	Contribution of the work	Application Area	Results Achieved
[26]	Jesus David et al. (2021)	Proposed DT combining IoT, real-time data and analytics to optimize crop productivity.	CEA	15-20% advancement in forecasted yield maximization (simulations report)
[27]	Mitsanis et al. (2024)	Launched 3D simulation framework of plant growth simulation used in testing of scenarios and resource planning.	Digital plant twins, simulation of plant growth	Not reported (modeling framework)
[28]	Kim & Heo (2024)	Crop monitoring of individual crops (irrigation, nutrients and yield).	Agriculture, high-tech farming.	Water usage reduction of up to 18 percent and yield per tree 12 percent improvement.
[29]	Dafang et al. (2023)	On-the-fly machine inspection, proactive maintenance, optimization of the way machines operate.	Smart mechanization, agriculture machinery	1015 percent decrease in the down time of machines; efficiency increased.
[30]	Tagarakis et al. (2024)	Overview of the applications of DT, advantages and future research.	Agriculture and forestry	Not reported

Source: Authors, (2026).

The comparison table 1 provides an overview of major works related to digital twins (DTs) in agriculture, including their contributions, areas of application, limitations, and outcomes. One study proposed a DT framework for controlled environment agriculture, where a 15-20% yield optimization is more applicable however, issues with scale and prohibitively expensive sensors remain. Another created a 3D plant simulation model to test scenarios, but was limited by the complexity of computations and crop validation. Mandarin-specific monitoring indicated 18% water consumption and 12% yield growth; however, these findings are not generalizable.

III. PROPOSED METHODOLOGY

The proposed methodology includes the creation of a Monitoring Digital Twin (mDT) to optimize the Controlled Environment Agriculture (CEA) processes with an integrated approach based on automation, real-time monitoring, and predictive analytics. Soil moisture, temperature, and humidity are also automatically controlled by a SELEC DIGIX-1 Programmable Logic Controller (PLC), which controls the conditions to the optimum growth conditions. The digital twin constantly gathers the information of multiple sensors located in the CEA setting and puts it in DynamoDB to analyze it in real-time and historically. Preprocessing of the collected data is done with a Median Filter with Z-Score Normalization to eliminate noise and normalize the data in order to have improved performance of the model. This is followed by extraction of significant attributes of the data after preprocessing through Gray Level Co-occurrence Matrix (GLCM), which contains information on space and texture that are important in crop health and yield. In the case of predictive analytics, a Deep Recurrent Q Network (DRQN) combined with a Convolutional Neural Network (CNN) is used to classify and predict the results, such as crop yield and the identification of weeds, pesticides, and plant diseases using the Crop Yield Prediction Dataset and Plant Village Dataset. The resulting information is rendered in the form of the SP112-GT40-S-CE Human Machine Interface (HMI), which provides operative information on such aspects as crop yield forecasting, weed identification, pesticide selection, and disease detection, both at the location and remotely.

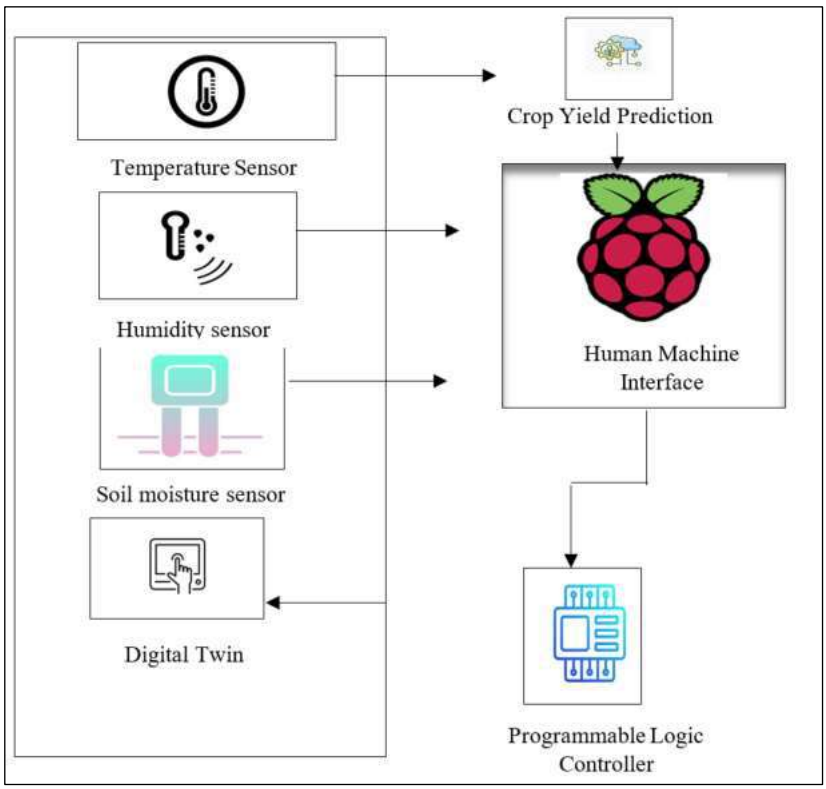


Figure 1: Proposed overall Block Diagram.
Source: Authors, (2026).

III.1 SMART ENVIRONMENT SETUP FOR DIGITAL TWIN-ENABLED CEA

The monitoring digital twin platform is set up on a CEA service, like a greenhouse or growth chamber setting. To ensure an uninterrupted monitoring of soil parameters (moisture, pH, and nutrient content), and climate parameters (temperature, humidity, CO2 concentration, and light density), the installation is prepared with a set of IoT-based sensors. A SELEC DIGIX-1 Programmable Logic Controller (PLC) is implemented to automate the functioning of the irrigation, ventilation, and lighting process on the basis of sensor inputs. Moreover, image sensors are used to monitor plant conditions, identifying weeds, pests, and diseases. An edge computing device connects the sensors and the PLC to process data on the edge, synchronize with a digital twin, and guarantee real-time monitoring and control. The machine learning system is similar with edges and has proven to be very efficient in real-time agricultural prediction and monitoring [31].

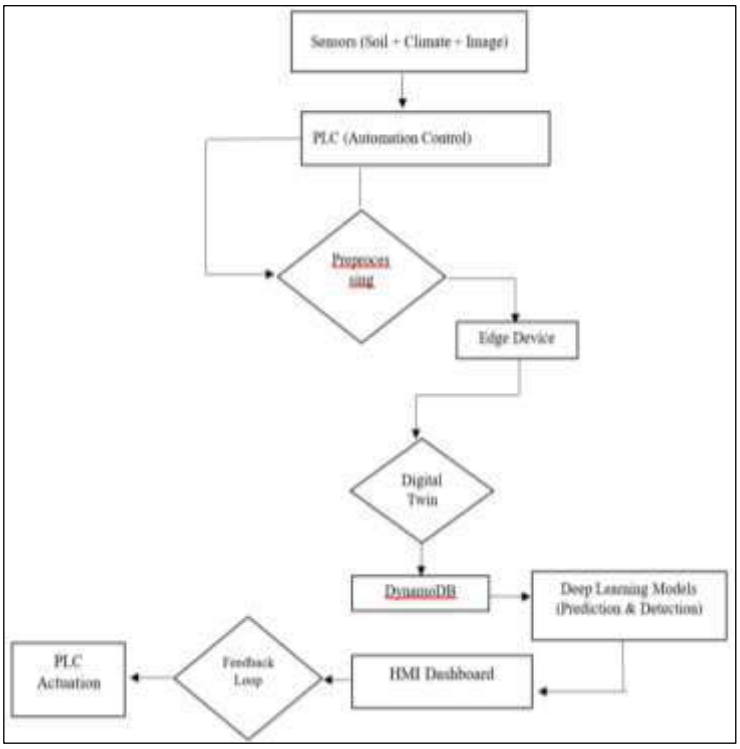


Figure 2: Environment Setup and Data Flow of the Proposed Monitoring Digital Twin (mDT) Framework.
Source: Authors, (2026).

Figure 2 represents the setup of the environment and data processing pipeline of the proposed mDT framework of CEA. Real-time data is measured by soil, climate, and image sensors and processed by a SELEC DIGIX-1 PLC to enable automation control. An edge device performs preprocessing at the location and aligns the data with the digital twin. The twin is linked to DynamoDB, which stores both historical and current data, allowing for prediction and detection through advanced analysis using deep learning models. Optimization of crop management is achieved by visualizing results using the HMI dashboard and implementing a feedback loop to keep the PLC updated.

III.2 PLANT VILLAGE DATASET

The Plant Village Dataset includes more than 54,000 images of leaf samples of 38 categories of healthy and diseased crops. The application is extensively applicable to training deep learning models, which can be used to accurately detect plant diseases, anomalies, and support smart agriculture to sustain farming practices. Sample leaf image provided in figure 3. Deep learning models such as ConvNeXt-Small have shown high accuracy in plant disease detection using leaf images, proving the reliability of CNN-based architectures for agricultural disease classification [32].

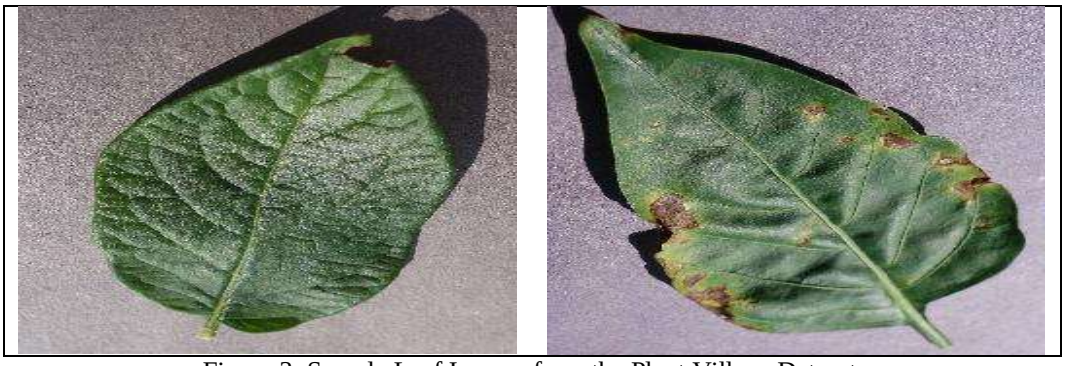


Figure 3: Sample Leaf Images from the Plant Village Dataset.
Source: Authors, (2026).

III.3 CROP YIELD PREDICTION DATASET

The science of training machines to learn and produce models for future predictions is widely used, and not for nothing. Agriculture plays a critical role in the global economy. With the continuing expansion of the human population understanding worldwide crop yield is central to addressing food security challenges and reducing the impacts of climate change. Crop yield prediction is an important agricultural problem. The Agricultural yield primarily depends on weather conditions (rain, temperature, etc), pesticides and accurate information about history of crop yield is an important thing for making decisions related to agricultural risk management and future predictions.

III.4 MEDIAN FILTER WITH Z-SCORE NORMALIZATION FOR PREPROCESSING

In the Edge Preprocessing layer, the raw sensor readings and crop images typically contain undesirable noise and deviations. A Median Filter is applied in order to enhance reliability in that impulsive noise like abrupt spikes or salt-and-pepper disturbances, is removed without distorting the necessary information. This makes the data smoother and more accurate for further processing. Z-score normalization is applied to standardize the data after noise has been removed, ensuring all features are on the same scale with equal variation. This hybrid method also enhances the quality of data and model learning, and yields superior prediction and detection accuracy within the system.

$$x'(t) = median\,\{x(t-k), \dots x(t), \dots., x(t+k) \tag{1}$$

Where the values are substituted with the median of the neighborhood surrounding the value to eliminate impulsive noise and leave details intact. Data is standardized with Z-score normalization after denoising.

$$x' = \frac{x-\mu}{\sigma} \tag{2}$$

The equation (2) features a zero mean and unit variance. This normalization ensures that there are no normalized and balanced inputs, allowing deep learning models to converge much more quickly and requiring a greater degree of accuracy, precision, and robustness in their prediction tasks.

III.5 DIGITAL TWIN LAYER

The Digital Twin layer is a virtual representation of the physical Controlled Environment Agriculture (CEA) system, synchronizing real-time data collected by soil, climate, and image sensors. It is connected to the SELEC DIGIX-1 PLC and edge device to reflect environmental conditions and operational states. DynamoDB stores both historical and current data, which can be leveraged to perform advanced analytics and identify anomalies. Times Twin supports predictive modeling including crop yield prediction and vegetation health evaluation, with deep learning algorithms. Digital Twin enhances decision-making, operational efficiency, and adaptive control in smart agriculture systems by providing a data-driven, dynamic simulation.

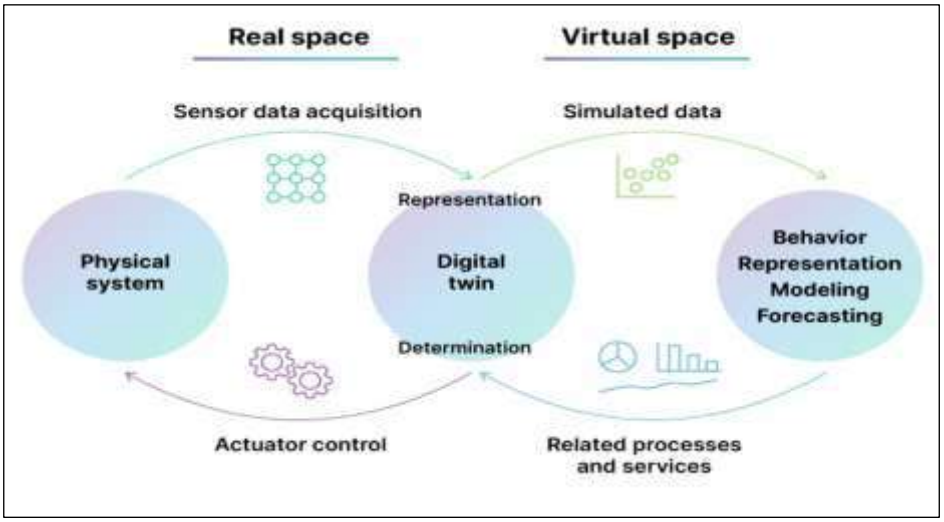


Figure 4: Digital Twin Layer Architecture in Smart Agriculture.
Source: Authors, (2026).

This figure 4 represents the Digital Twin layer in the mDT framework of Controlled Environment Agriculture (CEA). Data streams from soil, climate, and image sensors are fed into a virtual model in real time, which reflects the physical conditions. The Digital Twin is a system that links control systems and edge computing units to maintain constant synchronization and data storage. Based on this virtual representation, learned predictive analytics, such as crop yield forecasting and pest/disease identification through deep learning, are carried out. The model facilitates decision-making by allowing remote monitoring, simulation, and optimization of interventions, thereby increasing the efficiency and resilience of precision agriculture.

III.6 FEATURE EXTRACTION USING GRAY LEVEL CO-OCCURRENCE MATRIX

Gray Level Co-occurrence Matrix (GLCM) is a popular feature extraction method of studying image texture shown in figure 5. It analyses the frequency of occurrence of pairs of pixel intensities in a spatial relationship within an image, establishing a matrix that provides a representation of patterns of texture. Statistical features are generated out of this matrix, including contrast (degree of variation in intensities), energy (uniformity of texture), homogeneity (proximity of distribution to the diagonal), and correlation (linear dependence of gray levels).

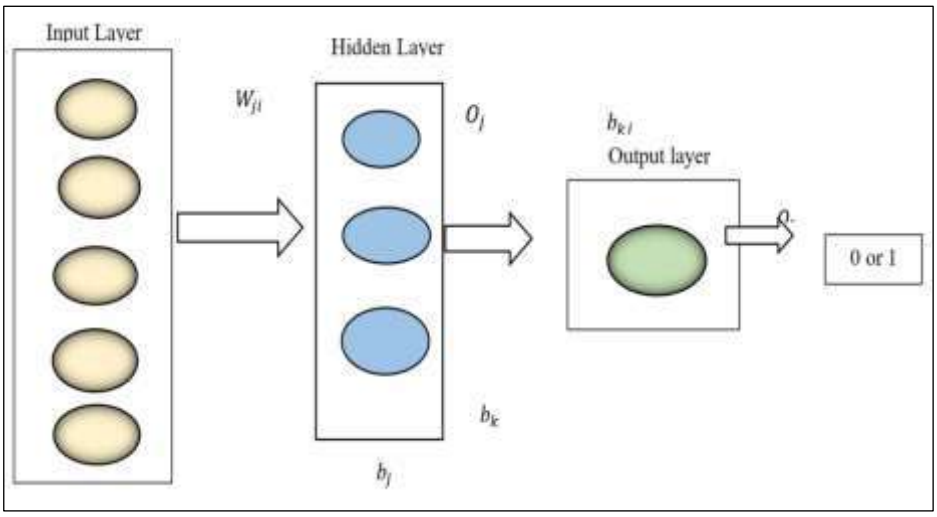


Figure 5: Gray Level Co-Occurrence Matrix Architecture.
Source: Authors, (2026).

III.7 CLASSIFICATION USING DEEP RECURRENT Q NETWORK MODEL WITH CONVOLUTIONAL NEURAL NETWORK (CNN)

The system proposed is based on a Deep Recurrent Q Network (DRQN) and a Convolutional Neural Network (CNN) in order to optimize the decision making and pattern recognition in crop monitoring. The CNN element isolates the spatial content of plant images, and disease pattern, growth of weeds, or pest infestation are highly recognized. The DRQN then processes these extracted features, adds temporal dependencies and sequence decision-making, to dynamic agricultural settings. The DRQN is learned through time unlike the static models and it adjusts to changes in the environment by learning the best control policies. This combination will guarantee proper detection of plant diseases, effective utilization of resources and the ability to make strong predictions regarding intelligent use of agriculture. ML-based predictive modeling studies also emphasize that optimized learning systems have a strong effect on improving the accuracy of prediction in a computational system [33].

Table 2: Algorithm Steps for DRQN with CNN.

Algorithm Steps for DRQN with CNN
Step 1: Preprocess and train plant images (including size, normalization, and augmentation) for training and testing.
Step 2: CNN layers (convolution, pooling, activation) extract spatial features in images.
Step 3: Feed CNN features into DRQN with recurrent layers to capture temporal dependencies.
Step 4: Apply Q-learning with epsilon-greedy policy to update action-value function based on rewards.
Step 5: Train model using experience replay and gradient descent to minimize Q-value prediction error.

Source: Authors, (2026).

The table 2 shows the sequential procedure of integrating CNN with DRQN to monitor agriculture. Step 1 is used to cleanse the input data and prepare it for training. The second step, step 2, isolates meaningful spatial content in images through CNN. Recurrent learning is incorporated in Steps 3 and 4 as part of Q-learning to make accurate decisions. Lastly, Step 5 focuses on training based on experience replay and optimization to enhance the performance of predictions.

IV. RESULTS AND DISCUSSION

The experimental analysis of the suggested mDT framework showed that both crop yield prediction and the anomaly detection tasks were significantly improved. DRQN-CNN model demonstrated better accuracy, precision, recall and F1-scores than the traditional methods, and low RMSE values of predictive consistency. Real-time sensor data combined with edge computing allowed making decisions quicker and minimizing latency to provide the opportunity to respond confidently to environmental changes. Moreover, the detection of weeds, pests and disease diagnosis became more reliable, which contributes to the active management of crops. Comprehensively, the findings indicate that the framework improves productivity, efficiency of resources, and sustainability of Controlled Environment Agriculture systems.

```
Using 'Year' as the rainfall target column.

Random sample index: 98
Predicted rainfall ('Year'): 2004.33

Feature values used for prediction:
Area Code          23
Element Code     5419
Item Code          27
Year Code        1972
Year             1972
Value           17213
Name: 3948, dtype: int64
```

Figure 6 Predicted Rainfall Using Crop Yield Prediction Dataset.
Source: Authors, (2026).

The figure 6 shows the machine learning prediction outcome on the Crop Yield Prediction Dataset. Some of the features used in the model include Area Code, Element Code, Item Code, Year Code, Year and Value, which are used to predict the rainfall. The input sample in this case is the Year 1972 with the value of 17213. Nonetheless, the model forecasts the year related to rainfall as about 2004.33 which implies that the target variable is the year column according to the trend of rainfall. This implies that this model is trying to infer the year using the past information on crop yield and rainfall, this information can be used to analyze the trend of climates and in agricultural planning.

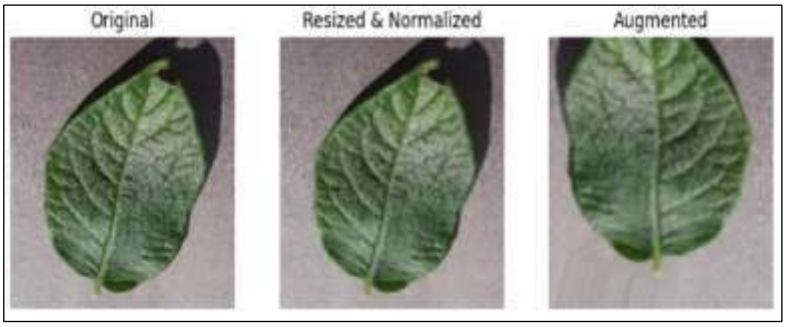


Figure 7: Image preprocessing for plant leaf analysis.
Source: Authors, (2026).

In agriculture, preprocessing is essential in the preparation of plant leaf images to undergo analysis in the form of deep learning. It starts with the original picture, which is reshaped to a specified size, making all the images in the dataset the same size. Normalization is then done to scale the pixel values between 0 and 1 that enables a faster convergence in the training. In order to enhance the robustness of the model even more, a set of data augmentation methods, including rotation, shifting, and flipping are proposed, producing various variants of the same picture. Such measures not only stop overfitting, but they also increase the capability of the model to identify plant diseases, pests and anomalies accurately in different environmental conditions depicted in figure 7.

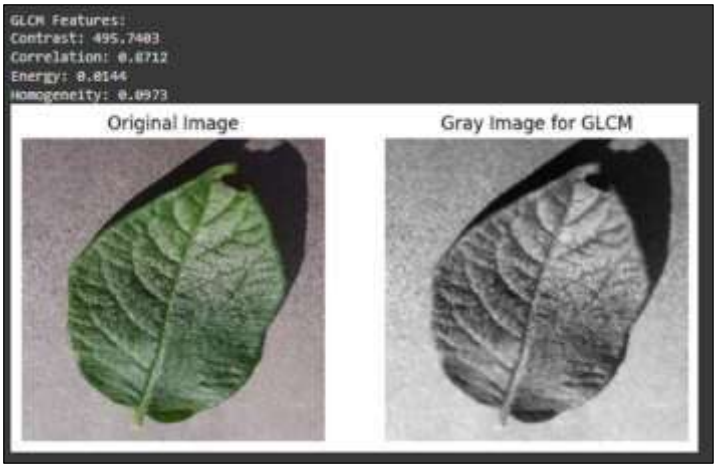
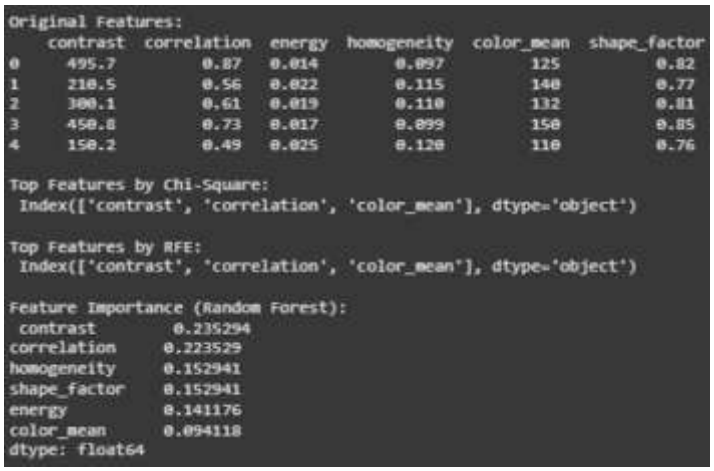


Figure 8: GLCM Texture Feature Extraction in the Analysis of Plant Leaf.
Source: Authors, (2026).

The diagram in figure 8 illustrates how the Gray Level Co-occurrence Matrix (GLCM) can be used to extract texture features out of a plant leaf. The original image is turned into grayscale and thus texture analysis is obtained by recording the relationship between pixel intensity. The features that are extracted are contrast, correlation, energy, and homogeneity that measure structural patterns in the leaf. These values are useful in the differentiation of healthy and diseased leaves by showing differences in texture. High contrast means high intensity variations, correlation measures pixel dependency, energy reflects uniformity, homogeneity takes care of smoothness. This type of feature extraction is important to precise disease detection and classification in precision agriculture



Original Features:

	contrast	correlation	energy	homogeneity	color_mean	shape_factor
0	495.7	0.87	0.014	0.097	125	0.82
1	210.5	0.56	0.022	0.115	140	0.77
2	300.1	0.61	0.019	0.110	132	0.81
3	450.8	0.73	0.017	0.099	150	0.85
4	150.2	0.49	0.025	0.120	110	0.76

Top Features by Chi-Square:
Index(['contrast', 'correlation', 'color_mean'], dtype='object')

Top Features by RFE:
Index(['contrast', 'correlation', 'color_mean'], dtype='object')

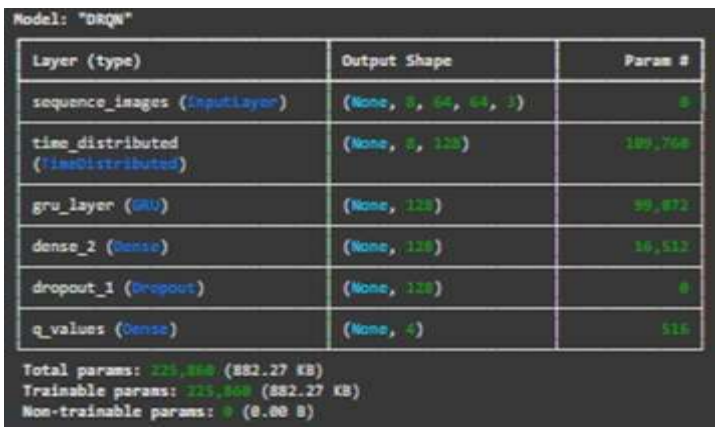
Feature Importance (Random Forest):

contrast	0.235294
correlation	0.223529
homogeneity	0.152941
shape_factor	0.152941
energy	0.141176
color_mean	0.094118

dtype: float64

Figure 9: Importance Analysis and Feature Selection of Plant Leaf.
Source: Authors, (2026).

The results of feature selection and ranking methods used on plant leaf data are presented in the figure 9. Original features like contrast, correlation, energy, homogeneity, colour mean and shape factor were taken into account. The most important feature common to both Chi-Square and RFE methods was contrast, correlation, and colour mean. This was also confirmed by the random Forest feature importance which ranked contrast and correlation as highly influential, then homogeneity and shape factor. The choice of those features brings out the key texture and colour patterns in leaves to detect and classify diseases more accurately, and compress dimensions to make the model more efficient and robust.



Model: "DRQN"

Layer (type)	Output Shape	Param #
sequence_images (InputLayer)	(None, 8, 64, 64, 3)	0
time_distributed (TimeDistributed)	(None, 8, 128)	189,760
gru_layer (GRU)	(None, 128)	99,072
dense_2 (Dense)	(None, 128)	16,512
dropout_1 (Dropout)	(None, 128)	0
q_values (Dense)	(None, 4)	516

Total params: 225,860 (882.27 KB)
Trainable params: 225,860 (882.27 KB)
Non-trainable params: 0 (0.00 B)

Figure 10: DRQN Model Overview of Sequential Image-based Decision Making.
Source: Authors, (2026).

DRQN model is a model of deep reinforcement learning that can process images in a sequence and decide them with reference to time patterns. It receives a series of 64×64×3 size images and it first processes them through Time Distributed layer to identify spatial features per frame. These attributes are then fed to a GRU layer that picks up the temporal dependencies and patterns in the sequence. Dense and dropout make the generalization and overfitting less significant. Lastly, the Q-values layer produces decision making action-value outputs. Having approximately 225K parameters, the model is good at balancing space and time-based learning on reinforcement tasks as demonstrated in figure 10.

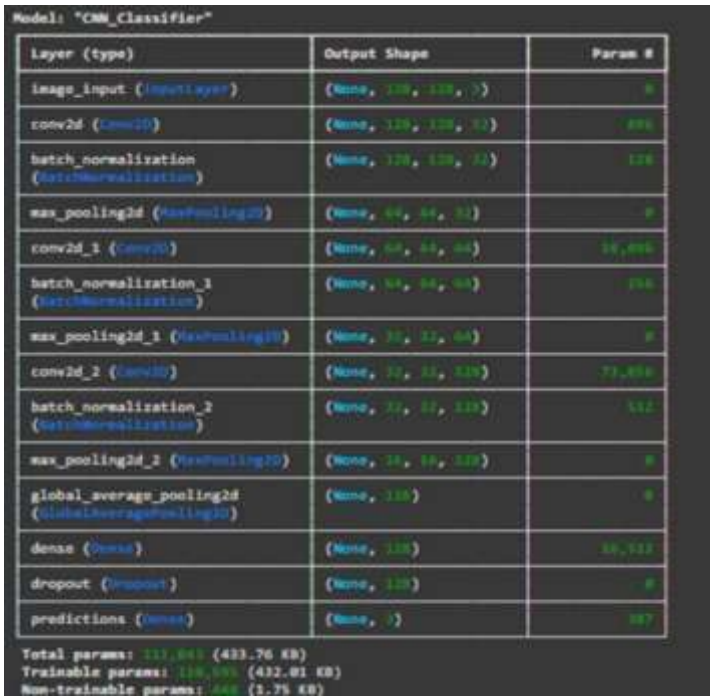


Figure 11: CNN Model Summary for Image Classification.
Source: Authors, (2026).

The CNN Classifier model shown is a lightweight deep learning model that is used in image classification. It takes input images of dimension 128x128x3 and processes them sequentially with convolutional layers which detect hierarchical features, including edges, textures and patterns. Max-pooling layers are used to reduce spatial dimensions and stabilize training by introducing batch normalization layers that preserve the main information, which is crucial, and ensure a cost reduction in computation. The global average pooling layer reduces feature map to a small representation and this is followed by dense and dropout layers to enhance generalization and decrease overfitting. Lastly, the prediction layer gives a probability of classifications. The model is effective in terms of efficiency and performance with about 111K parameters as illustrated in figure 10.



Figure 12: Training and Validation Accuracy and Loss per epoch.
Source: Authors, (2026).

Figure 12 illustrates the trend of training and validation accuracy (blue curves) and loss (red curves) with respect to the increasing number of epochs. The validation accuracy and training are gradual, the reason being that the model is learning well, and extrapolating the lesson to the unfamiliar material. At the same time, the training and validation loss decreases and decreases with a decrease in prediction errors. The close correlation between the training and validation curves also indicates that the model is neither overfitting nor underfitting, but that there is a compromise between memorizing the training data and achieving good performance on the validation data. This means a highly trained and healthy deep learning model.

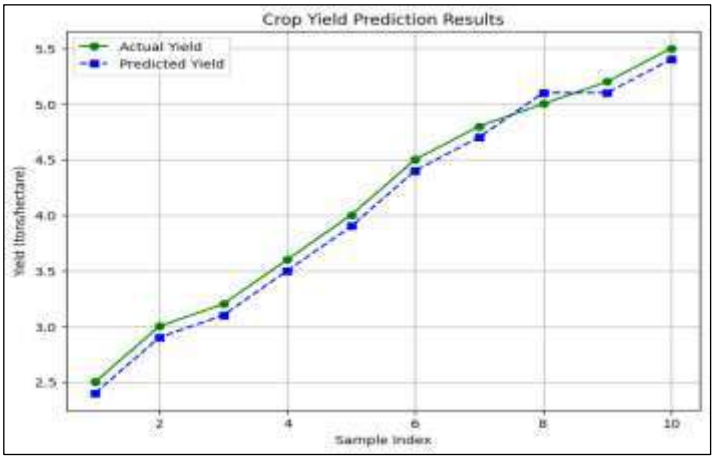


Figure 13: Crop Yield Prediction.
Source: Authors, (2026).

Figure 13 below is a comparison of the real crop yield values to the model estimates of the deep learning model. The observed yield curve or the green one reflects the observed yield and the blue dashed line reflects the predicted yield. The two curves are almost similar in trend, which means that the model is able to predict crop yield correctly in various samples. The small disparity between the actual and the anticipated values underscores the strength and the soundness of the method. This congruence indicates that the prediction model can help to make effective decisions in agriculture by giving relevant yield estimates, which translates to an improved resource planning process and sustainable farming.

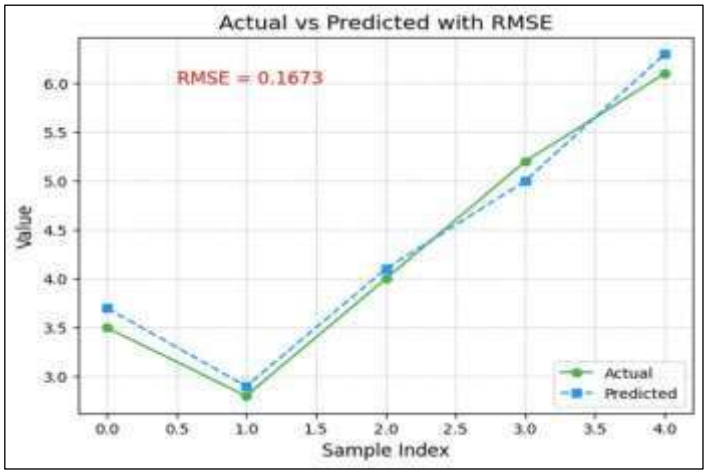


Figure 14: Actual vs Predicted with RMSE.
Source: Authors, (2026).

The figure 14 shows the comparison between actual and predicted values in the crop yield prediction model. These two lines are usually moving in the same direction where they show that the model is capable of providing results that are similar to the real-world outcomes. Root Mean Square Error (RMSE) of 0.1673 is low, and it implies that the average value of the difference between the predictive and actual values is low, which demonstrates the high precision of predictions. This minor mistake brings out the capability of the model to generalize well in different samples. This kind of forecasting is very important in the agricultural decision-making process as it allows farmers and other stakeholders to strategize resource distribution, minimize risk and enhance the overall production.

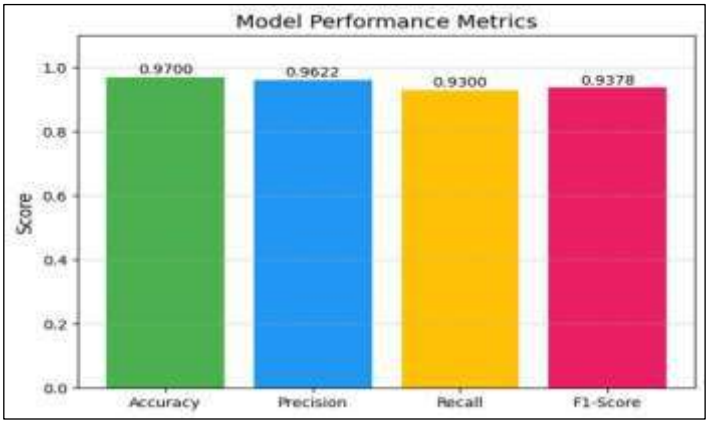


Figure 15: Performances Metrics.
Source: Authors, (2026).

The model presented in figure 15 was able to predict with an accuracy of 0.97% which indicates a high level of reliability. Its precision 0.962% points to the fact that it has reduced false positive rate, whereas its recall 0.93% points to its high sensitivity in detecting the relevant cases. The F1-score 0.937% is the combination between the precision and the recall, which ensures the overall effectiveness and soundness of the classification model.

V. CONCLUSION

This paper has proven that a Monitoring Digital Twin (mDT) framework is effective in Controlled Environment Agriculture (CEA), where deep learning, edge computing, and real-time automation can be used to maximize crop yield and achieve sustainable use of resources. Using SELEC DIGIX-1 PLC and other sensors to regulate the environment and to acquire data, the structure will correctly monitor and predict the crop yield and weed, pests, and plant diseases. The prediction and early detection of anomalies are further better than the conventional methods with the adoption of advanced deep learning models, specifically, Deep Recurrent Q Network with Convolutional Neural Networks. DynamoDB allows storage and retrieval of historical and real time data and the SP112-GT40-S-CE HMI provides the control to visualize that data and make informed decisions locally and remotely. The findings indicate that the application of digital twins with advanced AI models would greatly enhance productivity, efficiency, and sustainability of CEA systems in the future, as well as apply multi-crop environments, consider other environmental conditions like CO 2 concentration and humidity dynamics, and use adaptive control strategies through reinforcement learning to make autonomous decisions. In addition, cloud-edge hybrid computing can improve scalability and make data sharing between different farms possible, enabling the establishment of completely intelligent and self-optimizing smart farming systems.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: S. Sasikala and Sita Devi Bharatula.
Methodology: S. Sasikala and Sita Devi Bharatula.
Investigation: S. Sasikala and Sita Devi Bharatula.
Discussion of results: S. Sasikala and Sita Devi Bharatula.
Writing – Original Draft: S. Sasikala and Sita Devi Bharatula.
Writing – Review and Editing: S. Sasikala and Sita Devi Bharatula.
Supervision: S. Sasikala and Sita Devi Bharatula.
Approval of the final text: S. Sasikala and Sita Devi Bharatula.

VII. REFERENCES

[1] Liu, Zheng. "Intelligent monitoring and control of farmland based on edge-cloud collaboration and digital twin for digital energy management: investment benefit analysis." Sustainable Energy Research 12, no. 1, 2025.

[2] Kalyani, Yogeswaranathan, Nestor Velasco Bermeo, and Rem Collier. "Digital twin deployment for smart agriculture in Cloud-Fog-Edge infrastructure." International Journal of Parallel, Emergent and Distributed Systems 38, no. 6, 2023

[3] Peladarinos, Nikolaos, Dimitrios Piromalis, Vasileios Cheimaras, Efthymios Tserepas, Radu Adrian Munteanu, and Panagiotis Papageorgas. "Enhancing smart agriculture by implementing digital twins: A comprehensive review." Sensors 23, no. 16, 2023.

[4] Ghandar, Adam, Ayyaz Ahmed, Shahid Zulfiqar, Zhengchang Hua, Masatoshi Hanai, and Georgios Theodoropoulos. "A decision support system for urban agriculture using digital twin: A case study with aquaponics." IEEE Access 9, 2021.

[5] Rahman, Mohammad Abidur, Md Farhan Shahrior, Kamran Iqbal, and Ali Abushaiba. "Enabling Intelligent Industrial Automation: A Review of Machine Learning Applications with Digital Twin and Edge AI Integration." Automation 6, no. 3, 2025.

[6] Slimani, Hicham, Jamal El Mhamdi, and Abdelilah Jilbab. "Real-time greenhouse management using IoT, digital twin, and augmented reality for optimal control and decision-making." Acta IMEKO 14, no. 2, 2025.

[7] Jativa, Pablo Palacios, Ismael Soto, Cesar A. Azurdia-Meza, Iván Sánchez, Riu Wang, and Werther Kern. "Hybrid digital twin model for greenhouse and underground environments." IEEE Access 12, 2024.

[8] Singh, Rajmeet, Lakmal Seneviratne, and Irfan Hussain. "A deep learning-based approach to strawberry grasping using a telescopic-link differential drive mobile robot in ros-gazebo for greenhouse digital twin environments." IEEE Access, 2024.

[9] Lan, Hsun-Yu, Naomi A. Ubina, Shyi-Chyi Cheng, Shih-Syun Lin, and Cheng-Ting Huang. "Digital twin architecture evaluation for intelligent fish farm management using modified analytic hierarchy process." Applied Sciences 13, no. 1, 2022.

[10] Wang, Yuao, Jingjing Fang, Yao Cheng, Hao She, Yongan Guo, and Gan Zheng. "Cooperative end-edge-cloud computing and resource allocation for digital twin enabled 6g industrial iot." IEEE Journal of Selected Topics in Signal Processing 18, no. 1, 2023.

[11] Ali, Zain Anwar, Mahreen Zain, Raza Hasan, Hussain Al Salman, Bader Fahad Alkhamees, and Faisal Abdulaziz Almisned. "Circular economy advances with artificial intelligence and digital twin: Multiple-case study of Chinese industries in agriculture." Journal of the Knowledge Economy 16, no. 1, 2025.

[12] Jing Nie, Yi Wang, Yang Li and Xuewei Chao, "Artificial intelligence and digital twins in sustainable agriculture and forestry: a survey", Turkish Journal of Agriculture and Forestry, Vol. 46, No. 5, pp. 642-661, 2022.

[13] Orestis Spyrou, Mar Ariza-Sentís and Sergio Vélez, "Enhancing Education in Agriculture via XR-Based Digital Twins: A Novel Approach for the Next Generation", ASI, Volume 8, Issue 2, 2025.

[14] Aakash Thapa and Teerayut Horanont, "Digital twins in farming with the implementation of agricultural technologies", Applied Geography and Geoinformatics for Sustainable Development: Proceedings of ICGGS, pp. 121-132, 2022.

[15] Hsun-Yu Lan, Naomi A. Ubina, Shyi-Chyi Cheng, Shih-Syun Lin and Cheng-Ting Huang, “Digital Twin Architecture Evaluation for Intelligent Fish Farm Management Using Modified Analytic Hierarchy Process”, Applied Science, Vol. 13, No. 141, 2023.

[16] Yogeswaranathan Kalyani, Nestor Velasco Bermeo and Rem Collier, “Digital twin deployment for smart agriculture in Cloud-Fog-Edge infrastructure”, International Journal of Parallel, Emergent and Distributed Systems, Vol. 38, No. 6, pp. 461–476, 2023.

[17] Georg Goldenits, Kevin Mallinger, Sebastian Raubitzek, Thomas Neubauer, “Current applications and potential future directions of reinforcement learning-based Digital Twins in agriculture”, Smart Agricultural Technology, Vol. 8, 2024.

[18] Ahmad Ali Alzubi and Kalda Galyna, “Artificial Intelligence and Internet of Things for Sustainable Farming and Smart Agriculture”, IEEE Access, Vol. 11, 2023.

[19] Melanie Jans-Singh, Kathryn Leeming, Ruchi Choudhary and Mark Girolami, “Digital twin of an urban-integrated hydroponic farm”, Data-Centric Engineering, Vol. 1, No. e20, 2020.

[20] Fariza Sabrina, Shaleeza Sohail, Farnaz Farid, Sayka Jahan, Farhad Ahamed and Steven Gordon, "An Interpretable Artificial Intelligence Based Smart Agriculture System", Computers, Materials & Continua, vol.72, no.2, 2022.

[21] Cesco, Stefano, Paolo Sambo, Maurizio Borin, Bruno Basso, Guido Orzes, and Fabrizio Mazzetto. "Smart agriculture and digital twins: Applications and challenges in a vision of sustainability." European Journal of Agronomy, Vol.
146,2023.

[22] Nasirahmadi, Abozar, and Oliver Hensel. "Toward the next generation of digitalization in agriculture based on digital twin paradigm." Sensors 22, no. 2, 2022.

[23] Purcell, Warren, and Thomas Neubauer. "Digital Twins in Agriculture: A State-of-the-art review." Smart Agricultural Technology 3,2023.

[24] Singh, Maulshree, Rupal Srivastava, Evert Fuenmayor, Vladimir Kuts, Yuansong Qiao, Niall Murray, and Declan Devine. "Applications of digital twin across industries: A review." Applied Sciences 12, no. 11,2022.

[25] Peladarinos, Nikolaos, Dimitrios Piromalis, Vasileios Cheimaras, Efthymios Tserepas, Radu Adrian Munteanu, and Panagiotis Papageorgas. "Enhancing smart agriculture by implementing digital twins: A comprehensive review." Sensors, Vol. 23, no. 16, 2023.

[26] Jesus David, David Sanchez-Londono, and Giacomo Barbieri. "A digital twin architecture to optimize productivity within controlled environment agriculture." Applied Sciences, Vol. 11, no. 19, 2021.

[27] Mitsanis, Christos, William Hurst, and Bedir Tekinerdogan. "A 3D functional plant modelling framework for agricultural digital twins." Computers and Electronics in Agriculture, Vol. 218, 2024.

[28] Kim, Steven, and Seong Heo. "An agricultural digital twin for mandarins demonstrates the potential for individualized agriculture." Nature Communications, Vol. 15, no. 1 ,2024.

[29] Dafang,Yuefeng, W. U. Xiuheng, H. O. U. Siyu, L. I. Xiaoyu, Z. H. A. N. G. Yan'an "Digital twin for agricultural machinery: from concept to application." Smart Agriculture, Vol. 5, no. 2, 2023.

[30] Tagarakis, Aristotelis C., Lefteris Benos, George Kyriakarakos, Simon Pearson, Claus Gron Sorensen, and Dionysis Bochtis. "Digital twins in agriculture and forestry: a review." Sensors, Vol.24, no. 10,2024.

[31] B Seetharamulu, Kolli Maheedhar, Amit Sharma, Ram Charan, K Sarangam, B N Kumar Reddy, “An Efficient Predictive Alimentary System using Machine Learning on the edge,” 6th International Conference on Devices, Circuits and Systems
(ICDCS), 2022

[32] C. H. Lekkala, S. Bharatula, C. Sonika and K. Kannaiah, "ConvNeXt-Small: An Automated Deep Learning Method for Detecting Papaya Leaf Disease Using the BDPapayaLeaf Dataset," 2025 7th International Conference on Intelligent Sustainable Systems (ICISS), India, 2025, pp. 1369-1373, doi: 10.1109/ICISS63372.2025.11076525.

[33] B N K Reddy and Subrat Kar “Machine Learning Techniques for the Prediction of NoC Core Mapping Performance,” 26th IEEE Pacific Rim International Symposium on Dependable Computing (PRDC 2021), Dec 1– 4, in Perth, Australia, 2021.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 930-940. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3132

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

RANDOM FOREST GUIDED FEATURE WEIGHTED KNN MODEL FOR NUCLEAR REACTOR EVENT IDENTIFICATION

Arunprasath V[1], T V Santhosh[2], Gopika Vinod[3]

[1]Homi Bhabha National Institute, Mumbai, India
[2]Reactor Safety Division, Bhabha Atomic Research Centre, India
[3]Board of Research in Nuclear Sciences, India

[1]http://orcid.org/0009-0008-2401-4788, [2]http://orcid.org/0000-0003-3266-1324, [3]http://orcid.org/0000-0003-4460-7341

E-mail: *arunv@barc.gov.in

ARTICLE INFO

Article History
Received: December 13, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Machine learning,
Random Forest,
KNN,
Event detection,
Nuclear reactor monitoring.

ABSTRACT

Accurate identification of developing events is crucial in achieving reactor operation safety, decision making and accident management. Machine learning techniques are increasingly used as alternative to conventional model based methods due to their data driven nature for abnormal event detection. Among these, the K-Nearest Neighbors (KNN) algorithm remains widely used across industrial diagnostic systems, including applications in nuclear reactor monitoring, owing to its simplicity and good performance in multi class classification tasks. However, a key limitation of standard KNN is that all input features are treated with equal importance, which can limit its accuracy. To address this limitation, this study proposes a Random Forest guided feature weighted KNN classifier for reactor event identification. In the proposed approach, feature importance extracted from a Random Forest model is utilized to strengthen the ability of standard KNN model. The method is evaluated on a five class nuclear reactor event dataset and is found to outperform the base KNN and RF models. The performance of the proposed model is also compared with other KNN - RF ensemble methods and is shown to perform the best among all with superior performance metrics.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Early detection of abnormal conditions is very important to preserving reactor safety through timely intervention by operators. Conventionally, decisions about plant operation in the control room were based on alarms/trip systems, displays for monitoring plant conditions, and emergency operating procedures. These systems have limitations because they were developed from simplifications of fundamental physical laws and they operate in response to those events which are covered under such principles. Further, accurate modelling of nuclear reactors via first principles could be challenging due to the complicated process behavior, therefore it limits the uses of model-based diagnostics entirely. On the other hand, in contrast to first principle methods, data driven methods learn patterns from operational data and real time measurements and hence do not require detailed mathematical descriptions regarding reactor behavior. Since modern industrial plants including nuclear plants acquire large amounts of operational data, the information can be utilized to train machine leaning models to identify developing events in their earlier stages. Whenever an event occurs, operators in a nuclear reactor control room is expected to analyze many signals, alarms and trends simultaneously under huge mental pressure. Human error has been identified as an important contributor to nuclear accidents worldwide. Hence, machine learning techniques can be used to develop data driven systems learning from huge volumes of operational data to provide diagnostic support to the operators reducing human errors. Owing to their benefits in terms of safe and cost effective reactor operation, many techniques for the identification and diagnosis of events have been studied. In general, such approaches aim to represent relationships within reactor systems or derive informative features through pattern recognition strategies, enabling the detection of abnormal reactor conditions based on measured process data [1]. In recent years, there has been a growing shift toward the use of data driven methodologies for monitoring nuclear reactors.

Currently, a wide range of machine learning techniques—such as neural networks, support vector machines (SVMs), dimensionality reduction algorithms, ensemble learning frameworks, regression models, and others—have been adopted to forecast the operational behavior of nuclear power plants, as noted by Hu G et al. [2]. Moreover, machine learning tools, including advanced deep learning architectures, are increasingly applied for condition monitoring, equipment health evaluation, and remaining useful life (RUL) estimation within nuclear facilities [3], [4]. Studies illustrate the use of SVM based models [5], [6], where deviations in process variables are exploited to identify potential accident scenarios in reactors. Reference [7] proposes a diagnostic approach that not only detects faults but also quantifies their severity by combining Principal Component Analysis with neural network techniques. In [8], Santosh et al. developed a neural network driven framework to diagnose severe coolant loss events in nuclear power plants. Additional research has explored a variety of data driven solutions for assessing and managing plant wide and component level health conditions, as summarized in [9].

Reference [10] examines multiple classification techniques—such as PCA, SVM, decision trees, and hybrid schemes—for identifying equipment and actuator faults in PWR systems. Meanwhile, [11] reports that ensembles based on nearest neighbor classifiers achieved the highest accuracy for fault detection, whereas ensembles composed solely of decision trees performed comparatively poorly. Motivated by these works, our earlier work involved development of KNN model for the purpose of transient event identification in reactors [12]. The developed KNN model achieved an accuracy of 97% which is a hold out test data performance. In this work, a novel method is proposed for an improved KNN model for event identification with enhanced accuracy. The proposed model is a Random Forest guided feature weighted KNN classifier to improve the accuracy of classification. By integrating the local feature relevance derived from Random Forests with the neighbourhood based decision structure of KNN, the proposed model is shown to outperform the basic KNN model and other KNN-RF ensemble models. Different performance metrics of the model such as categorical accuracy, confusion matrix, class wise performance score are studied and are shown to provide the best results when compared with the other models.

II. PROBLEM STATEMENT AND DATASET

II.1 EVENT DETECTION

Systems that support timely event detection are crucial to maintaining the safe operation of any nuclear reactor. Research or test reactors, in particular, consist of highly intricate subsystems designed both for operational safety and for carrying out diverse experimental activities. This paper focuses on identifying abnormal events in such test reactors using a data driven strategy built upon a novel machine learning framework. In these facilities, the reactor core is usually located within a large pool filled with demineralized light water, which serves simultaneously as coolant and moderator. Under normal operating conditions with offsite electrical power available, primary cooling pumps ensure forced circulation around the core. Heated coolant is routed through heat exchangers, where its thermal energy is transferred to the secondary cooling loop. The secondary loop employs its own set of pumps that drive water toward cooling towers for ultimate heat removal.

If offsite power is lost, core cooling relies on natural circulation within the pool, enabled by passive valves engineered to open automatically and permit flow. Throughout operation, vital plant variables—including process parameters and equipment status indicators—are continuously monitored and archived. Because of their extensive instrumentation, these reactors generate and log large volumes of process data [13], [14]. Within nuclear reactor safety design and assessment, Postulated Initiating Events (PIEs) play a critical role. These events represent potential triggers that may drive the reactor away from its normal operational state, possibly progressing into anticipated operational occurrences or even accident scenarios. Rapid identification and mitigation of such events are essential, making the precision of detection systems or models highly important. For this reason, the present study approaches the challenge through the application of an appropriate and innovative machine learning methodology.

II.2 DATASET

Training effective machine learning models relies on the availability of credible and well defined datasets. In this work, the training data are generated by simulating key PIE scenarios using thermal hydraulic codes capable of reproducing the dynamic behavior of accident conditions in light water reactors, including research reactors [15], [16]. Such events represent transients which are departures from normal conditions that may occur due to abnormal system behavior. These transients can stem from a variety of sources, including equipment malfunctions, operator mistakes, and both internal and external hazards. For the reactor examined in this study, the relevant events were selected in accordance with IAEA SSR-3, the international standard that outlines essential safety requirements for research reactors [17]. Four significant events typical of test research reactors were chosen, and simulated data from these scenarios were used for model development. Table 1 summarizes the set of events considered in this work. Each scenario is generated through a detailed event sequence analysis, which produces time dependent trends for the plant parameters affected during the transient. For the four selected events, 16 operational variables were identified as influential. These 16 measurements serve as the feature set used to train and construct the machine learning models. Consequently, the dataset is a labeled one comprising five classes, including a baseline class that represents normal operation with no event. The feature set includes both continuous variables such as reactor power, coolant flow rate, and heat exchanger temperatures, and categorical indicators, including various reactor trip signals, alarm activations, and the status of passive cooling valves, among others.

Table 1: Events modeled for detection.

S.No.	Event
1	Failure of offsite power event
2	Failure of pump flow event
3	Loss of coolant event
4	Failure of heat sink event

Source: Authors, (2026).

III. METHODOLOGY

III.1 K-NEAREST NEIGHBORS CLASSIFIER

One of the simplest and most understandable supervised learning methods for classification is the K-Nearest Neighbors (KNN) algorithm. The main principle is that in the feature space, data points belonging to the same class typically appear near one another. Therefore, we may use the labels of the closest samples in the training set to estimate the class of a new, unseen sample. Considering training dataset D,

$$\{(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)\}, \tag{1}$$

where $x_1 \epsilon \ \mathbb{R}^d$ represents a d dimensional feature vector and y_i is the corresponding class label, the goal is to assign class to a new sample x. The KNN classifier performs this in the following steps:

1. Measure the distance between the test sample x and all samples in the training set.
2. Identify the nearest K samples, referred to as the "neighbors."
3. Perform majority voting among these Kneighbors.
4. The class receiving the highest number of votes is assigned to x.

Mathematically, this can be expressed as:

$$\hat{y}(x) = mode\{y(1), y(2), \dots, y(K)\}, \tag{2}$$

where $y(i)$ denotes the class label of the i-th nerest neighbor.

Distance metric plays a crucial role in KNN. Commonly used metrics include:

- Euclidean distance

$$\mathrm{d_E(x, x_i)} = \sqrt{\sum_{\mathrm{j=1}}^{\mathrm{d}} \left(\mathrm{x_j - x_{i,j}}\right)^2} \tag{3}$$

- Manhattan distance

$$d_M(x, x_i) = \sum_{j=1}^{d} \left| x_j - x_{i,j} \right| \tag{4}$$

The decision boundary formed by KNN is non linear and adapts to the data distribution. This characteristic makes it suitable for complex classification tasks such as nuclear event identification, where the feature relationships are often non linear. The choice of Ksignificantly affects the classifier's performance. A small K makes the classifier sensitive to noise and outliers, potentially overfitting. A larger K smooths the decision boundary and improves generalization but may overlook small scale structures in the data [18]. KNN is a relatively simple to use and straightforward algorithm. Since there is no training phase, it is adaptable to incremental data. KNN is one of the most suitable models to handle multiclass problems. However, KNN suffers from the curse of dimensionality and is sensitive to duplicate or noisy features [19]. Without specific weighting, contribution of each feature is treated equally by the algorithm. These limitations are especially pertinent to complicated industrial datasets, such nuclear reactor transients, where some features could be more important than others. This insight naturally emphasizes the necessity for an improved classifier that guides the decision making process by incorporating extra knowledge, such as feature importance information.

III.2 RANDOM FOREST CLASSIFIER

Random Forest is an ensemble learning method that constructs a collection of decision trees and aggregates their outputs to improve classification reliability. The core idea is that multiple weak learners, when combined in an intelligent manner, can form a strong and robust classifier. Random Forests are particularly effective for complex, nonlinear datasets such as those encountered in nuclear reactor monitoring, where different features may influence the event dynamics in varying ways [20]. The concepts of random feature selection and bootstrap aggregation (bagging) are used to construct a Random Forest. A bootstrap sample taken from the original dataset is used to train each tree in the forest. Furthermore, only a randomly chosen subset of features is taken into account for determining the optimal split during each tree's growth. Because the trees are diverse due to these two types of randomness, there is less chance of associated errors [21]. Let the training dataset be

$$\{(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)\} \tag{5}$$

Where $x_1 \epsilon \ \mathbb{R}^d$ represents a d dimensional feature vector and y_i is the corresponding class label. For each tree T_b, a bootstrap sample $D_b$is drawn by sampling N times with replacement from the original dataset D. Some samples may appear multiple times, while others may be excluded (out of bag samples). At each internal node of tree T_b, instead of evaluating all features, only a randomly selected subset $F_b \subseteq \{1,2, \dots, d\}$ is considered for splitting. Because of this, there is reduced correlation between trees and it helps capture different structural relationships in the data. Each decision node chooses the feature and threshold that offers the greatest reduction in impurity. One of the most used impurity measure, Gini impurity is defined for a dataset D as:

$$Gini(D) = 1 - \sum_{k=1}^{c} p_k^2 \tag{6}$$

Where p_k is the proportion of samples belonging to class k, and C is the total number of classes. If a split on feature f divides the data into subsets D_{left} and D_{right}, the impurity reduction is

$$\Delta Gini(f) = Gini(D) - \left\{ \frac{|D_{left}|}{|D|} Gini(D_{left}) + \frac{|D_{left}|}{|D|} Gini(D_{right}) \right\} \tag{7}$$

The feature and threshold giving the maximum ΔGini are selected for that node [22]. Once all trees are trained, classification of a new sample x is performed by passing it through each tree. Each tree produces an output $T_b(x)$, and the forest prediction is obtained through majority voting:

$$\hat{y}(x) = mode\{T_1(x), T_2(x), \ldots, T_3(x)\} \tag{8}$$

Where B is the total number of trees. The misclassifications of each tree are often uncorrelated because they are trained on distinct subsets of data and characteristics. The ensemble greatly lowers variance and enhances generalization when combined. The capacity of Random Forests to produce a quantitative assessment of feature relevance is a key feature. The impurity reduction ascribed to each feature across all trees is averaged to calculate this.

III.3 RF–KNN HYBRID LEARNING APPROACHES

Random Forest and KNN possess complementary characteristics. While Random Forest is a powerful ensemble learner capable of handling nonlinear relationships and providing feature relevance information, KNN is a simple yet effective local classifier that operates based on distance relationships. Several hybrid strategies have been proposed in literature to utilize the strengths of both models. The most commonly adopted approaches are summarized below.

III.3.1 Voting Based RF–KNN Ensemble

In voting based ensembles, RF and KNN act as independent base classifiers. For a given test sample x, each classifier produces a class prediction, and the final output is determined based on majority voting [23]. If $\hat{y}_{\text{RF}}(x)$ and $\hat{y}_{\text{KNN}}(x)$ represent the predictions from RF and KNN respectively, the ensemble decision is given by

$$\hat{y}(x) = mode\,\{\hat{y}_{\text{RF}}(x), \hat{y}_{\text{KNN}}(x)\} \tag{9}$$

In soft voting, the class probabilities are averaged:

$$P(c \mid x) = \frac{1}{2}\left(P_{RF}(c \mid x) + P_{RF}(c \mid x)\right) \tag{10}$$

Although this approach often improves robustness, it does not exploit the internal knowledge of either model beyond their final predictions.

III.3.2 Stacking Based RF–KNN Ensemble

In stacking, RF and KNN serve as base learners, while a separate meta classifier is trained to combine their outputs [24]. Let the predictions of RF and KNN for a sample x be concatenated as:

$$z(x) = [\hat{y}_{\text{RF}}(x), \hat{y}_{\text{KNN}}(x)] \tag{11}$$

A meta classifier $f_m(\cdot)$ is then trained such that:

$$\hat{y}(x) = f_m(x) \tag{12}$$

Stacking enables the learning of nonlinear relationships between base classifier outputs, but it introduces additional model complexity and it requires careful cross validation in order to avoid overfitting.

III.3.3 Feature Level Fusion of RF and KNN

In feature level fusion, transformed feature representations obtained from RF are concatenated with the original input features and given to a KNN classifier or another learning model [23]. If x is the original feature vector and $h_{\text{RF}}(x)$ represents the RF induced feature transformation, the fused feature vector is:

$$x_{fusion} = [x, h_{\text{RF}}(x)] \tag{13}$$

KNN is then applied in this expanded feature space. This approach increases representational power but also increases dimensionality and computational cost. While the above strategies combine RF and KNN at the decision level or feature level, they do not explicitly utilize the feature importance knowledge generated by RF to guide the distance computation in KNN, which forms the central idea of the proposed model.

III.3.4 Proposed RF Guided Feature Weighted KNN Model

The proposed model integrates Random Forest and KNN in a more informed manner by embedding feature importance knowledge from RF directly into the KNN distance metric. Unlike conventional fusion strategies, here RF does not merely act as an independent classifier but provides guidance to KNN in the form of feature relevance. In standard KNN, all features contribute equally to the distance computation. However, in nuclear reactor systems, different parameters contribute unequally to various transient events. Noisy or weakly relevant features can distort distance measurements and lead to incorrect neighbor selection. Therefore, introducing feature weight awareness into KNN is essential to improve classification reliability. Random Forest offers a natural framework for estimating feature relevance based on impurity reduction [25]. In this work, this information is exploited to construct a feature weighted distance metric for KNN.

III.3.4.1 Feature Importance Extraction using Random Forest

A Random Forest consisting of Bdecision trees is trained on D. For each feature f_j, the importance is computed by aggregating the Gini impurity reduction over all the trees:

$$I(f_j) = \frac{1}{B}\sum_{b=1}^{B}\sum_{n \in T_{b(f_j)}} \Delta Gini_{b,n} \tag{14}$$

where $\Delta\text{Gini}_{b,n}$ denotes the impurity reduction achieved by node n of tree T_b when splitting on feature f_j. The resulting importance vector is:

$$\boldsymbol{I} = [I(f_1), I(f_2), \ldots, I(f_d)] \tag{15}$$

This vector is normalized such that:

$$w_j = \frac{I(f_j)}{\sum_{k=1}^{d} I(f_k)} \tag{16}$$

The normalized vector

$$W = [w_1, w_2, \ldots, w_d] \tag{17}$$

Represents the relative importance of each feature.

III.3.4.2 Feature Weighted Distance Metric for KNN

The conventional Euclidean distance between two samples x and $x_i$is defined in equation 3. In the proposed model, this is modified into a feature weighted Euclidean distance:

$$d_w(x, x_i) = \sqrt{\sum_{j=1}^{d} w_j (x_j - x_{i,j})^2} \tag{18}$$

Where w_j represents the importance weight of the j-th feature obtained from Random Forest. This formulation ensures that features with higher diagnostic relevance have a stronger influence on the neighbor selection process, while less informative features are automatically suppressed.

III.3.4.3 RF Guided KNN Classification Procedure

For a test sample x:

1. Compute weighted distances $d_w(x, x_i)$ to all training samples.
2. Identify the Knearest neighbors based on these distances.
3. Assign x to the class receiving the majority vote among its neighbors:

$$\hat{y}(x) = mode\{y_{(1)}, y_{(2)}, \ldots, y_{(k)}\} \tag{19}$$

The operational sequence of the model is given in figure 1. This structure highlights how Random Forest acts as a knowledge provider, guiding the neighbor selection process of KNN.

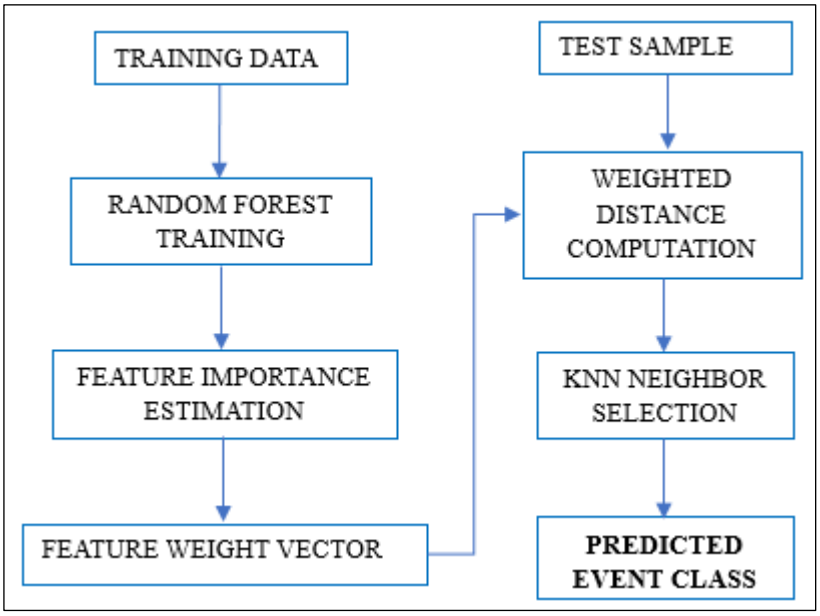


Figure 1: Operational sequence of the proposed RF guided feature weighted KNN model.
Source: Authors, (2026).

III.3.5 Data Preprocessing and Modeling Strategy

III.3.5.1 Data Preprocessing and preparation

All input features are normalized using z-score normalization to eliminate scale dominance:

$$x' = \frac{x-\mu}{\sigma} \tag{20}$$

where μ and σ denote the mean and standard deviation of a feature respectively.
The dataset is divided into three mutually exclusive subsets:

- **Training set** – used to train the Random Forest and KNN
- **Validation set** – used for hyperparameter optimization
- **Test set** – used for final performance evaluation

This separation ensures unbiased performance estimation.

III.3.5.2 Modeling and Hyperparameter Configuration

The performance of KNN is highly sensitive to the choice of K and the distance metric. To obtain the optimal configuration, a grid search combined with cross validation is employed on the training validation set.

The following hyperparameter ranges are explored:

- $K \in \{1,2,\ldots,30\}$
- Distance metric $\in$ {Euclidean, Manhattan}

For each parameter combination, cross validation accuracy is computed, and the configuration that yields the maximum average accuracy is selected. The optimal hyperparameters obtained for this study are: $K = 27, Distance\ Metric =$ Euclidean. The Random Forest model is trained to extract feature importance information required for the proposed RF guided weighted KNN framework. Optimal hyperparameters are obtained after performing grid search combined with cross validation. The following hyperparameter ranges are explored:

- No. of trees $N \in \{10,30,\ldots,200\}$

Other parameters of the model are chosen as

- Splitting criterion : Gini
- Maximum number of features at each split : $\sqrt{d}$, where d is the number of features.
- Maximum tree depth : None
- Minimum samples to split an internal node: 2
- Minimum samples in a leaf node: 1

The optimal number of trees obtained from hyperparameter tuning is N =70. After completing the independent modeling stages of KNN and Random Forest, both models are integrated by incorporating the feature importance weights in to the Euclidean distance calculation for the KNN classifier. Since the proposed RF guided weighting modifies only the distance metric and not the neighborhood structure, the optimal value K = 27, obtained from the standard KNN optimization, was retained for the weighted KNN model. After hyperparameter selection, the final RF guided weighted KNN model is trained using the combined training and validation data. The trained model is then evaluated on an independent hold out test dataset to assess its generalization capability.

IV. PERFORMANCE METRICS

The performance of the proposed RF guided feature weighted KNN classifier is evaluated using several standard classification metrics, namely accuracy, confusion matrix, precision, recall, and F1-score [26]. Since the considered problem involves multi class nuclear reactor event identification, it is essential to assess not only the overall classification accuracy but also the class wise prediction behavior. Each metric used in this study is briefly described below. The confusion matrix is the fundamental tool used to visualize the classification performance. For a multi class problem with C classes, the confusion matrix is a $C \times C$ matrix, where the element M_{ij} denotes the number of samples belonging to class i that are predicted as class j. For a single class k, the following quantities are defined:

- **True Positives (TP)**: samples correctly classified as class k
- **False Positives (FP)**: samples incorrectly classified as class k
- **False Negatives (FN)**: samples of class k misclassified as other classes
- **True Negatives (TN)**: samples correctly classified as not belonging to class k

The confusion matrix provides detailed insight into which events are correctly detected and which are confused with one another. A typical confusion matrix for a 2 class problem is shown below:

$$\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix} \tag{21}$$

Accuracy represents the proportion of total correctly classified samples to the total number of samples:

$$Accuracy = \sum_{i=1}^{C} M_{ii} \tag{22}$$

where:

- M_{ii} represents the diagonal elements of the confusion matrix (correct classifications),
- C is the number of classes,
- N is the total number of samples.

Although accuracy offers a broad indicator of performance, it does not show how effectively individual classes are recognized, particularly in situations when class distributions are unbalanced. For safety critical applications like nuclear reactor event identification, precision alone is therefore insufficient. Precision quantifies the percentage of samples that are correctly predicted to belong to a particular class. For class k:

$$Precision_k = \frac{TP_k}{TP_k + FP_k} \tag{23}$$

Precision is particularly important when false alarms are costly. In reactor monitoring, high precision ensures that when a specific transient is predicted, the likelihood of it being a false detection is low. Recall (also known as sensitivity or true positive rate) measures how many of the actual samples of a class are correctly identified:

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \tag{24}$$

In safety related systems, recall is crucial since it shows how well the model can identify real anomalous events. A low recall suggests that the system might miss some crucial transients, which could result in dangerous situations. The F1-score is a fair performance indicator for every class and is the harmonic mean of precision and recall. F1 score is defined as:

$$F1_k = 2 * \frac{Precision_k * Recall_k}{Precision_k + Recall_k} \tag{25}$$

V. RESULTS AND DISCUSSIONS

This section presents the experimental results obtained for the standard KNN classifier, Random Forest classifier, various RF–KNN hybrid ensemble methods, and the proposed RF guided feature weighted KNN model. Initially, the individual base models were trained and optimized independently. Subsequently, different ensemble strategies, namely voting ensemble, stacking ensemble, and feature level fusion, were implemented and evaluated.

Finally, the proposed RF guided feature weighted KNN model was trained using the optimized hyperparameters and tested on a hold out dataset. The results are analyzed using accuracy trends, feature importance visualization, comparative performance plots, and the normalized confusion matrix. Figure 1 illustrates the variation of KNN classification accuracy as a function of the number of neighbors K. It can be observed that the accuracy increases rapidly for small values of K, indicating improved generalization as noise sensitivity reduces. Beyond this region, the performance gradually stabilizes as K increases. The accuracy reaches a peak and remains nearly constant in the range $K = 25$ *to* $K = 30$. indicating that the classifier has achieved a good balance between bias and variance. Based on this trend, $K = 27$ is selected as the optimal value for the KNN classifier and is used for all the hybrid models. Figure 2 shows the classification accuracy of the Random Forest classifier for different tree counts. As the number of trees increases, the performance first improves because to better ensemble averaging and reduced variation. However, adding more trees doesn't really improve the situation until the accuracy hits saturation. The plot clearly shows that the best accuracy is obtained with 70 trees, and the accuracy stabilizes after roughly 100 trees. Beyond this, raising the ensemble size primarily leads to increased computing costs and overfitting. This shows that a moderate number of trees is sufficient for precise feature extraction and classification in the given situation.

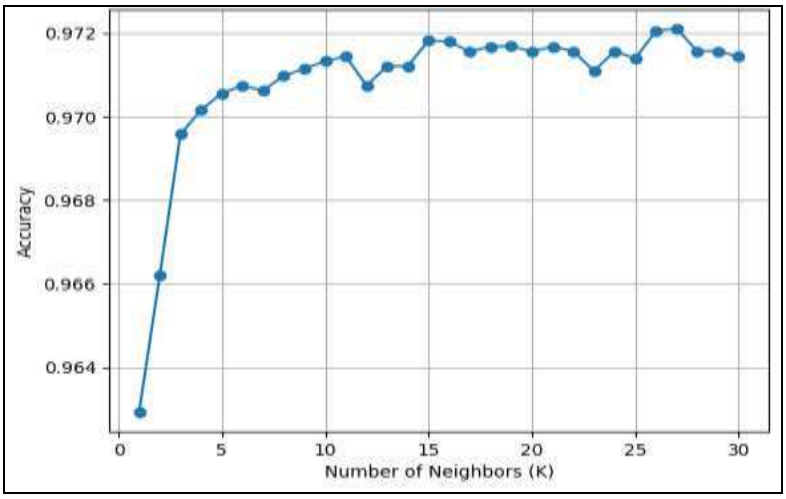


Figure 1: Accuracy Vs K (No. of Neighbors) plot of KNN classifier model.
Source: Authors, (2026).

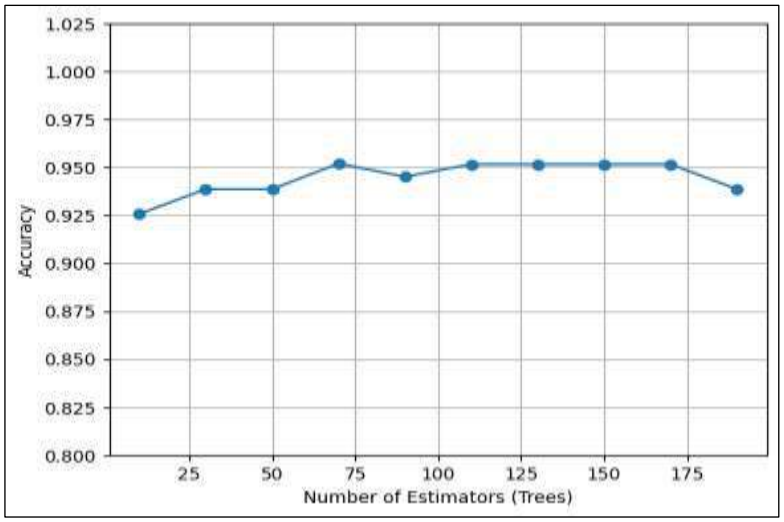


Figure 2: Accuracy Vs No. of estimators (trees) plot of Random Forest classifier model.
Source: Authors, (2026).

The trained Random Forest model's feature importance ratings are shown in Figure 3. It is evident that just a small number of features have a significant influence on the classification choice, whereas a number of features have comparatively little significance. The basic rationale for the suggested method—that considering all features equally in distance computation is not optimal for nuclear reactor event classification—is clearly supported by this observation. In the next KNN classification stage, the availability of such important information offers a solid foundation for the introduction of feature weighted distance computation. Performance metrics for several models are shown in Table 2, including the individual class F1 score and the total classification score for 1) Standard KNN 2) Random Forest 3) stacking ensemble 4) Voting ensemble 5) Feature level fusion 6) Proposed RF guided feature weighted KNN model. An overview of the models' comparative categorization accuracy is given in Figure 4. Strong baseline performance is demonstrated by the standard KNN classifier's high accuracy of 97.07%. At about 95.16%, the Random Forest classifier's accuracy is marginally lower. Although they offer a slight improvement over Random Forest, the stacking and voting ensembles are still not as good as the conventional KNN. Compared to the classical ensembles, the feature level fusion model produces a slight improvement. Nevertheless, the suggested RF guided feature weighted KNN model outperforms all individual models and hybrid baseline models with the maximum accuracy of 98.57%.

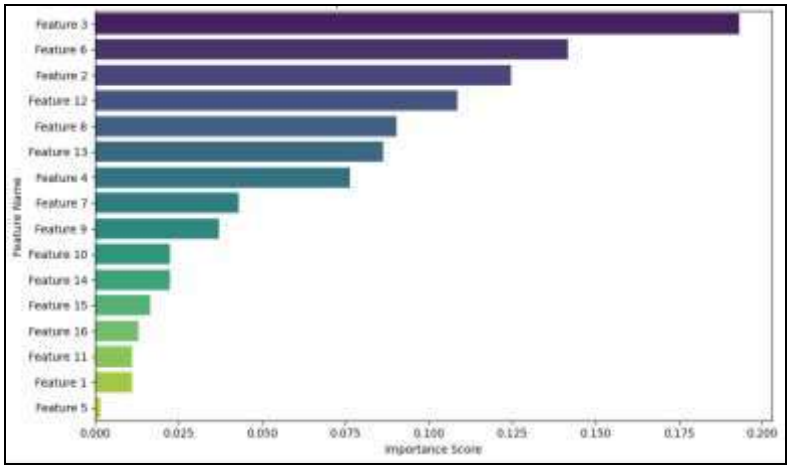


Figure 3: Feature importance plot of Random Forest Classifier model.
Source: Authors, (2026).

Table 2: Comparison of performance metrics of different models.

Model	Categorical accuracy	F1-Score				
		Class 1	Class 2	Class 3	Class 4	Class 5
KNN	97.07%	0.9	1	1	1	0.92
Random Forest	95.16%	0.88	1	1	1	0.91
KNN-RF Stack ensemble	95.86%	0.9	1	1	1	0.92
KNN-RF Voting ensemble	95.19%	0.88	1	1	1	0.91
KNN-RF feature level fusion	95.90%	0.89	1	1	1	0.92
Proposed RF guided feature weighted KNN model	98.57 %	0.96	1	1	1	0.97

Source: Authors, (2026).

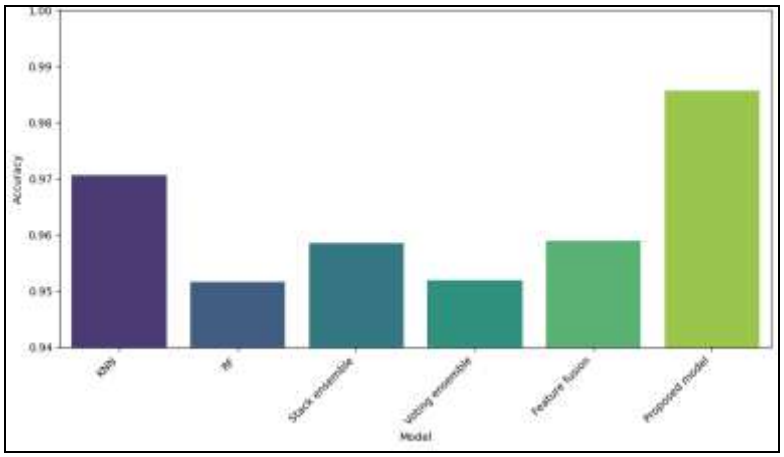


Figure 4: Classification accuracy of different models.
Source: Authors, (2026).

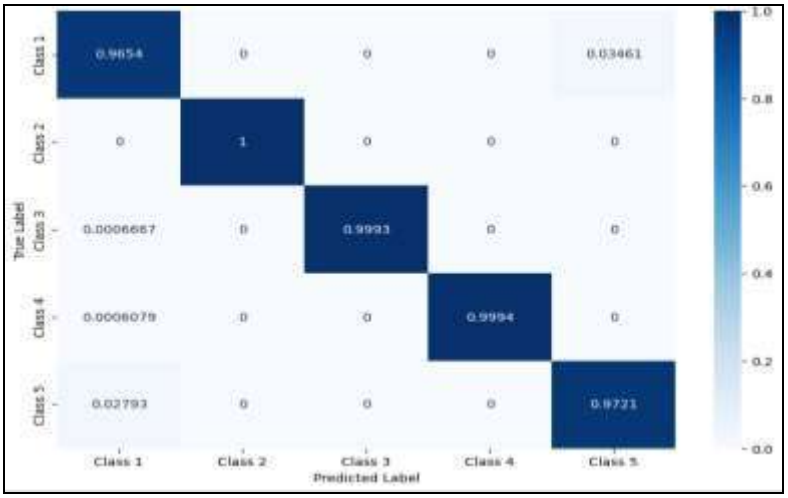


Figure 5: Normalized confusion matrix of the proposed RF guided feature weighted KNN model.
Source: Authors, (2026).

The proposed approach yields substantial improvements in the F1-scores of the relatively challenging event categories, namely Class 1 (No event) and Class 5 (Failure of heat sink event), where the F1-score improves from approximately 0.88–0.92 to 0.96 and 0.97, respectively. The remaining classes exhibit near perfect F1-scores across all models, indicating strong class separability. These improvements are consistent with the normalized confusion matrix in figure 5, which shows pronounced diagonal dominance and significantly reduced misclassification for the proposed model. This confirms that directly embedding Random Forest derived feature relevance into the KNN distance metric is more effective than combining classifiers only at the decision or feature fusion level.

V. CONCLUSIONS

In this work, a Random Forest guided feature weighted KNN classifier was developed to improve the performance of multi class reactor event classification. The key idea behind the proposed model is the integration of feature importance knowledge derived from Random Forest into the distance computation of KNN, thereby enabling the classifier to emphasize diagnostically significant reactor parameters while suppressing the influence of less informative features. The individual performances of KNN and Random Forest were first evaluated, followed by the implementation of several hybrid RF–KNN ensemble strategies, including stacking, voting, and feature level fusion. While these conventional ensembles offered marginal improvements, none of them surpassed the baseline KNN performance significantly. The proposed model, however, demonstrated a clear performance advantage. With the optimized hyperparameters and feature weighted distance formulation, the proposed RF guided KNN model achieved an overall classification accuracy of 98.57%, outperforming all baseline and ensemble models.

Class wise F1-scores further confirmed the robustness of the proposed method, especially for the more challenging event categories, where the F1-scores improved substantially compared to other approaches. The normalized confusion matrix illustrated strong diagonal dominance and minimal misclassification, validating the improved detection capability of the proposed model. These results collectively demonstrate that directly embedding Random Forest feature relevance into the KNN decision framework is a more effective strategy than combining the two models only at the decision or fusion stage. Overall, the proposed RF guided feature weighted KNN approach offers a simple and highly accurate method for nuclear reactor event identification. With improved performance, the work contributes to the growing field of safety in reactor operation and operator support, particularly where dependable event detection is critical.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Arunprasath V, T V Santhosh, Gopika Vinod.
Methodology: Arunprasath V, T V Santhosh, Gopika Vinod.
Investigation: Arunprasath V, T V Santhosh, Gopika Vinod.
Discussion of results: Arunprasath V, T V Santhosh, Gopika Vinod.
Writing – Original Draft: Arunprasath V, T V Santhosh, Gopika Vinod.
Writing – Review and Editing: Arunprasath V, T V Santhosh, Gopika Vinod.
Supervision: Arunprasath V, T V Santhosh, Gopika Vinod.
Approval of the final text: Arunprasath V, T V Santhosh, Gopika Vinod.

VII. ACKNOWLEDGMENTS

The authors greatly thank Mr.P.K.Guchhait, (Bhabha Atomic Research Centre, India) for his support in simulation and data generation.

VIII. REFERENCES

[1] Uhrig, R. E., Tsoukalas, L. H. "Soft computing technologies in nuclear engineering applications", Progress in Nuclear Energy, 34 (1) pp. 13–75, 1999. DOI:10.1016/S0149-1970(97)00109-1.

[2] Hu, G., Zhou, T., Liu, Q. " Data-Driven Machine Learning for Fault Detection and Diagnosis in Nuclear Power Plants: A Review", Frontiers in Energy Research, 9 (663296),2021. DOI: 10.3389/fenrg.2021.663296

[3] Yong X., Yunze C., Lin S., "Condition Assessment of Nuclear Power Plant Equipment Based on Machine Learning Methods: A Review", Nuclear Technology, 209(7), 929-962,2023. DOI:10.1080/00295450.2023.2169042

[4] Wang, F., Liu, A., Qu, C., Xiong, R., Chen, L., "A Deep-Learning Method for Remaining Useful Life Prediction of Power Machinery via Dual-Attention Mechanism", Sensors, 25(2), 497, 2025. DOI:10.3390/s25020497

[5] Gottlieb, C., Arzhanov, V., Gudowski W., et al. "Feasibility Study on Transient Identification in Nuclear Power Plants Using Support Vector Machines", Nuclear Technology, 155(1), pp. 67-77, 2006. DOI:10.13182/NT06-A3746

[6] Author,2024.

[7] Saeed, H.A., Peng, M.J.,Wang, H., et al. "Novel fault diagnosis scheme utilizing deep learning networks", Progress in Nuclear Energy, 103066,2020. DOI: 10.1016/j.pnucene.2019.103066

[8] Tian, D., Deng, J., Colantuono, G., Vinod, G., Santosh, T.V., & Gorse, C., "Identification of Loss of Coolant Accidents of Nuclear Power Plants Using Artificial Neural Networks", Nuclear Power Plant Life Management, Vienna, International Atomic Energy Agency (IAEA), 2021. DOI: 10.1016/j.ress.2008.08.005

[9] Gomez-Fernandez, M., Higley, K., Tokuhiro, A., Welter, K., Wong, W., & Yang, H., "Status of research and development of learning-based approaches in nuclear science and engineering: A review", Nuclear Engineering and Design, 359,
110479, 2020. https://doi.org/10.1016/j.nucengdes.2019.110479

[10] Ben Q., Jingang L., Jiejuan T., “Fault Diagnosis Techniques for Nuclear Power Plants: A Review from the Artificial Intelligence Perspective”, Energies, 16,1850,2023. https://doi.org/10.3390/en16041850

[11] Naimi, A., Deng, J., Doney, P., Sheikh-Akbari, A., Shimjith, S.R. and Arul, A.J., “Machine learning-based fault diagnosis for a PWR nuclear power plant”, IEEE Access, 10, pp.126001-126010, 2022.
https://doi.org/10.1109/ACCESS.2022.3225966

[12] Author, 2023.

[13] Mauro Cappelli, Instrumentation and Control Systems for Nuclear Power Plants, ed. 1, Woodhead Publishing Series in Energy, ed. 1, Mauro Cappelli, Woodhead Publishing, Cambridge,2023.

[14] IAEA, Modern Instrumentation and Control for Nuclear Power Plants: A Guidebook, Technical Reports Series No. 387, ed 1, IAEA, International Atomic Energy Agency, Vienna,1999.

[15] INL, RELAP5/Mod3.2 code manual, Vol. I-VII, Report NUREG/CR-5535, U.S. NRC, Washington D.C,1995.

[16] Guo, Y., Wang, G., et al., Thermal hydraulic analysis of loss of flow accident in the JRR-3M research reactor under the flow blockage transient.,Annals of Nuclear Energy, 118,147-153,2018. https://ui.adsabs.harvard.edu/link_gateway/2018AnNuE.118..147G/doi:10.1016/j.anucene.2018.04.014

[17] IAEA, Safety of research reactors, IAEA Specific Safety Requirements SSR-3, ed 1, International Atomic Energy Agency, Vienna,2016.

[18] Yang, L., Jin, R.: ‘Distance metric learning: A comprehensive survey’, Technical report, Department of Computer Science and Engineering, Michigan State University, 2007.

[19] I. A. Abu Amra and A. Y. A. Maghari, "Students performance prediction using KNN and Naïve Bayesian," 2017 8th International Conference on Information Technology (ICIT), Amman, Jordan, 2017, pp. 909-913, doi: 10.1109/ICITECH.2017.8079967

[20] Parmar, A., Katariya, R., Patel, V. (2019). A Review on Random Forest: An Ensemble Classifier. In: Hemanth, J., Fernando, X., Lafata, P., Baig, Z. (eds) International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018. ICICI 2018. Lecture Notes on Data Engineering and Communications Technologies, vol 26. Springer, Cham.,2018. https://doi.org/10.1007/978-3-030-03146-6_86

[21] Daho, Mostafa El Habib, Chikh, Mohammed Amine, Journal of Medical Imaging and Health Informatics, Volume 5, Number 3, pp. 539-544(6), June 2015.

[22] Tangirala S. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. International Journal of Advanced Computer Science and Applications,11(2):612-9,Feb 2020. 10.14569/ijacsa.2020.0110277.

[23] Rojarath, Artittayapron, Wararat Songpan, and Chakrit Pong-Inwong. "Improved ensemble learning for classification techniques based on majority voting." In 2016 7th IEEE international conference on software engineering and service science (ICSESS), pp. 107-110. IEEE, 2016. 10.1109/ICSESS.2016.7883026.

[24] Mienye ID, Sun Y. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. Ieee Access, Sep 16;10:99129-49,2022. 10.1109/ACCESS.2022.3207287.

[25] Rogers, Jeremy, and Steve Gunn. "Identifying feature relevance using a random forest." In International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection", pp. 173-184. Berlin,
Heidelberg: Springer Berlin Heidelberg, 2005. https://doi.org/10.1007/11752790_12

[26] Vujović, Željko. "Classification model evaluation metrics." International Journal of Advanced Computer Science and Applications 12, no. 6: 599-606,2021. 10.14569/IJACSA.2021.0120670.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 941-950. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3137

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

SMART PLS STRUCTURAL MODELLING OF AI-DRIVEN HR PRACTICES, ENTREPRENEURIAL SUCCESS AND EMPLOYER REPUTATION IN INDIAN FIRMS

Nidhi Sonkar[1], Amit Kumar Pandey[2], Ratchagaraja Dhairiyasamy[3] Deekshant Varshney[4], Subhav Singh[5]

[1]Assistant Professor, Institute of Management Studies, Ghaziabad (U.P), India.
[2]Professor, Institute of Management Studies, Ghaziabad (U.P), India.
[3]Department of Electronics and Communication Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai – 602105, Tamil Nadu, India.
[4]Centre for Research Impact and Outcome, Chitkara University, Rajpura – 140417, Punjab, India.
[5]Chitkara Centre for Research and Development, Chitkara University, Himachal Pradesh – 174103, India.

[1]https://orcid.org/0000-0002-0151-4295, [2]https://orcid.org/0000-0001-5767-6333, [3]https://orcid.org/0000-0001-7528-7585, [4]https://orcid.org/0000-0003-3485-0284, [5]https://orcid.org/0009-0009-0586-6034

E-mail: ratchagaraja@gmail.com, nidhisonkar16@gmail.com, amitkbp@gmail.com, profvarshney@outlook.com, er.subhav@hotmail.com

ARTICLE INFO

Article History
Received: December 14, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Artificial intelligence in human resource management,
Partial least squares structural equation modelling,
Entrepreneurial achievement,
Employer reputation,
AI system quality

ABSTRACT

Artificial intelligence is transforming human resource management by automating recruitment, employee development and performance management in entrepreneurial firms. Evidence on mechanisms linking AI-enabled HR practices, entrepreneurial achievement and employer reputation has been limited. This study aims to quantify relationships among AI adoption, AI system quality, entrepreneurial achievement and company reputation. Survey data from 110 HR professionals were analysed using SmartPLS with a five-point Likert scale. AI adoption affected perceived AI quality ($\beta = 0.769$, $P < 0.001$), indicating that integrating AI into HRM was associated with dependable systems. AI quality affected entrepreneurial achievement ($\beta = 0.815$, $P < 0.001$), entrepreneurial achievement significantly influenced company reputation ($\beta = 0.375$, $P = 0.033$), and direct links from AI adoption to entrepreneurial achievement and company reputation were not significant ($\beta = 0.027$, $P = 0.352$; $\beta = 0.030$, $P = 0.226$). These findings indicate that reputational gains stem from deploying AI in HR processes, and future research will examine longitudinal dynamics of AI-enabled HRM to support SDG-oriented governance.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Artificial intelligence (AI) is becoming a foundational capability in contemporary organizations, reshaping decision processes, coordination mechanisms, and the use of information across functions. Human resource management (HRM) stands at the center of this transformation because people, skills, and learning determine whether technological innovation translates into sustained competitive advantage. HR functions now incorporate algorithmic screening, predictive analytics, and adaptive learning systems, which influence workforce composition, capability development, and internal perceptions of fairness and opportunity [1]. For entrepreneurial firms facing intense uncertainty and resource constraints, the quality of decisions on recruitment, employee development, and performance management is especially critical, making AI-enabled HR practices a strategic concern rather than a peripheral efficiency tool [2]. HRM encompasses a set of practices that attract, select, develop, evaluate, and retain employees in line with organizational goals. AI-based solutions already support candidate sourcing, résumé parsing, job–applicant fit matching algorithms, adaptive training platforms, and continuous performance monitoring. Studies on AI in organizational contexts reported gains in information processing, reduced cycle times, support for more evidence-based HR decisions, and improvements in coordination between HR and line managers [3].

Prior research on AI in management information systems and organizational performance examined these tools as part of broader digital infrastructures and often treated HR as one subsystem among many rather than as a distinct strategic domain. In parallel, research on entrepreneurial performance examined the roles of innovation capabilities, dynamic capabilities, and learning orientation, and it frequently highlighted talent and knowledge as central resources in young firms [4]. Entrepreneurial firms operate under volatile conditions where opportunities arise and fade quickly, customer needs change rapidly, and new business models appear alongside new technologies. In such settings, HRM is not only administrative; it shapes the capacity to experiment, scale promising initiatives, and pivot in response to feedback. AI-enabled HR practices promise better matching between roles and competencies, faster identification of potential, and more tailored development pathways. Earlier studies on AI in HRM examined discrete elements of this promise, such as automated recruitment screening, AI-supported learning platforms, and digital tools for performance reviews [5]. These contributions highlighted efficiency gains and sometimes reported positive responses from HR professionals and employees, yet they seldom connected those micro-level improvements to entrepreneurial outcomes such as innovation, adaptability, or new venture performance.

Another strand of work focused on reputational consequences of AI use in organizations. Research on employer branding and corporate reputation suggested that technology-intensive HR practices influence external perceptions of a firm as modern, innovative, and data-driven. Some studies reported that the visible use of AI in recruitment and employee development enhanced perceptions of professionalism and fairness among applicants and employees, while others raised concerns about opacity, bias, and privacy. Ethical debates around transparency, accountability, and bias mitigation in AI-based decision systems highlighted risks to organizational legitimacy when stakeholders questioned the foundations of automated judgments [6]. Despite this attention, prior work mostly examined large established firms or sector-specific cases and rarely linked AI-enabled HR practices in entrepreneurial firms to broader measures of company reputation as perceived by external stakeholders and potential employees. Within the entrepreneurship literature, researchers have long argued that digital technologies support opportunity discovery, resource orchestration, and scaling. AI-driven analytics in marketing, operations, and finance were studied as enablers of better forecasting, more accurate customer segmentation, and improved risk assessment. Empirical work on entrepreneurial firms that adopted AI often emphasized gains in operational efficiency and innovation, measuring outcomes through financial performance, productivity, or innovation outputs.

Yet these studies rarely considered HR as the primary locus of AI integration, even though hiring decisions, skill development, and performance management directly condition the firm’s capacity to exploit AI in other functions. Furthermore, entrepreneurial achievement was usually conceptualized through broad indicators without explicit reference to the contribution of AI-enabled HR practices [7]. A further limitation of prior research concerned the distinction between AI adoption and AI quality. Much of the empirical work treated adoption as a binary or scalar variable, focusing on whether organizations used AI tools and to what extent. Less attention was paid to attributes such as reliability, accuracy, predictive capability, and user trust, even though these attributes determine whether AI-based recommendations translate into effective decisions. Conceptual discussions suggested that poorly configured or low-quality AI systems could undermine HR outcomes by amplifying bias or generating misleading signals, whereas high-quality systems could augment human judgment. A small number of studies proposed that the quality of AI implementation could mediate the relationship between adoption and performance outcomes, but systematic empirical tests of this proposition remained limited, particularly in HR settings within entrepreneurial firms [8]. From a methodological perspective, many empirical contributions relied on single-path regression models or qualitative case studies, which constrained the ability to examine complex relationships among adoption, system quality, entrepreneurial achievement, and reputation.

Although structural equation modeling (SEM) is well-suited to analyzing latent constructs and mediated relationships, relatively few studies have used this approach to examine AI-enabled HR practices in entrepreneurial contexts. Those who adopted SEM often focused on technology acceptance or user intention rather than on entrepreneurial outcomes and company reputation [9]. As a result, existing evidence did not clarify the pathways through which AI adoption in HR, the quality of AI systems, and entrepreneurial achievement interact to shape perceptions of the firm as an attractive and innovative employer. The context of emerging economies adds another layer of importance to this topic. Entrepreneurial firms in countries such as India face distinctive constraints in talent availability, infrastructure, and institutional support, yet they also facetrong pressures to modernize and adopt digital tools [10]. Earlier studies on AI integration in HRM and entrepreneurial performance drew heavily on data from advanced economies or specific sectors, limiting the generalizability of their findings to settings in which firms operate with different resource profiles and regulatory environments. Evidence from Indian entrepreneurial firms that have already adopted AI-based HR systems remains sparse, especially when research requires access to HR professionals' internal perceptions and assessments of AI-enabled practices. This gap leaves practitioners without clear empirical guidance on which dimensions of AI use in HR matter most for entrepreneurial performance and reputation building in such settings [11]. Collectively prior research left several questions unresolved. The field lacks quantitative evidence that connects AI adoption in HRM, the perceived quality of AI systems, entrepreneurial achievement, and company reputation within a single integrated framework.

Existing work seldom treated entrepreneurial achievement as a mediating mechanism linking AI-related constructs to external perceptions of the firm, even though theory suggests that reputational benefits arise when stakeholders associate the organization with successful innovation and adaptability. Furthermore, validated measurement scales for AI adoption and AI quality in HR practices within entrepreneurial firms have not been widely applied, limiting comparability across studies and weakening the empirical basis for managerial recommendations [12]. The present study addresses these limitations by focusing on entrepreneurial firms that have integrated AI into core HR practices, specifically recruitment, employee development, and performance management, and by modeling the relationships among AI adoption, AI quality, entrepreneurial achievement, and company reputation using a structural equation approach with Smart Partial Least Squares. The study introduces a framework in which AI adoption and AI quality function as distinct but related constructs, entrepreneurial achievement serves as a mediating variable, and company reputation serves as the key outcome from the perspective of employer image. By focusing on HR-focused AI integration in entrepreneurial firms in India, the study addresses the need for context-sensitive evidence on technology-enabled HRM and its strategic consequences [13]. The purpose of this study is to examine the extent to which AI adoption and AI system quality in HR practices contribute to entrepreneurial achievement in entrepreneurial firms, and to assess the role of entrepreneurial achievement in shaping company reputation in an emerging economy context using Smart PLS

II. MATERIALS AND METHODS

The study was designed as a quantitative, cross-sectional investigation of the integration of Artificial Intelligence into Human Resource practices within entrepreneurial firms in India. A structured online questionnaire was administered to Human Resource professionals, and the collected data were subjected to variance-based structural equation modelling using SmartPLS. The methodological process extended from construct operationalization and instrument development to sampling and data collection, and then to data preparation, measurement model evaluation, and structural model estimation. The sequential stages of the research design are presented in Figure 1, and the principal constructs and indicators used in the measurement model are summarised in Table 1. The target population consisted of Human Resource professionals employed in entrepreneurial firms operating in India. Entrepreneurial firms were defined as organizations founded within the past 15 years, demonstrating sustained growth in employment or revenue, and actively pursuing innovation in products, services, or processes. Inclusion criteria required respondents to occupy decision-making or specialist roles in Human Resource Management, such as HR manager, talent acquisition lead, learning and development specialist, or HR business partner, and to possess at least two years of professional experience in HR.

Firms of medium to large size were prioritised, as these organizations were assumed to have sufficient digital infrastructure to support the adoption of Artificial Intelligence-enabled HR tools. A non-probability sampling strategy was applied. Professional contact networks, industry associations, and business networks were used to identify potential respondents. Invitation messages containing a brief study description and a link to the online questionnaire were distributed via email and professional networking platforms. Data were collected exclusively through the online survey, and no paper-based responses were accepted. A total of 110 valid responses from HR professionals working in entrepreneurial firms were obtained and retained for analysis, consistent with sample size recommendations for partial least squares structural equation modelling with the number of indicators and paths specified in the conceptual framework. The sampling and data collection parameters are summarised in Table 2. Before full deployment, the questionnaire was pretested with a small group of HR professionals and researchers in the domain of Human Resource Management and digital transformation. Feedback from this pretest was used to refine item wording, sequence, and layout, with particular attention given to clarity, terminological consistency, and avoidance of ambiguous phrasing.

The final instrument consisted of three sections. The first section recorded demographic and professional information, including age, gender, highest educational qualification, years of HR experience, current role, organizational size, and sector. The second section captured information on the extent and nature of Artificial Intelligence usage in HR-related processes within the respondent's organization. The third section contained the psychometric scales used to operationalize the latent constructs in the conceptual model [14]. All latent constructs were operationalized as reflective variables, measured on a 5-point Likert-type scale. Respondents were instructed to indicate the extent of agreement with each statement on a scale from 1 ("strongly disagree") to 5 ("strongly agree"). Artificial Intelligence adoption captured the extent to which AI-enabled tools and applications were employed in core HR functions, including recruitment, employee development, and performance management. Artificial Intelligence quality reflected respondents' perceptions of the reliability, accuracy, and usefulness of deployed AI systems in HR contexts. Entrepreneurial achievement captured the contribution of AI-supported HR practices to innovation, operational efficiency, and attainment of entrepreneurial goals. Company reputation measured perceptions of the organization as an attractive, innovative, and sustainable employer in the labour market and among external stakeholders. Illustrative items for each construct are presented in Table 1.

Table 1. Constructs, abbreviations, and illustrative measurement items.

Construct	Abbreviation	Illustrative item (5-point agreement scale)
Artificial Intelligence adoption	AA	"Use of AI technology in recruitment processes was widespread in the organization."
Artificial Intelligence quality	AQ	"AI systems were capable of predicting HR needs accurately."
Entrepreneurial achievement	EA	"Use of AI in HR contributed to achieving the entrepreneurial goals of the organisation."
Company reputation	CR	"The organisation was perceived as an attractive employer by potential and current employees."

Source: Authors, (2026).

Instrument layout and visual design were implemented using a commercially available online survey platform. A consistent font, spacing, and colour scheme were applied across all pages to reduce visual fatigue. Items belonging to different constructs were interspersed to reduce response patterning. Mandatory response fields were used for central psychometric items, while sensitive demographic questions included an explicit "prefer not to answer" option. Skip logic was activated to route respondents who did not meet the inclusion criteria away from the main questionnaire. Time stamps recorded the start and end of each session to identify responses with unrealistically short completion times [15]. Data collection was conducted continuously, with the survey link remaining active at all times. Reminder messages were issued at spaced intervals to individuals who had not yet responded. Each respondent was provided with a unique survey link to prevent duplicate submissions. Before accessing the questionnaire, respondents were presented with an electronic information sheet outlining the research objectives, data-handling procedures, and assurances of anonymity. Informed consent was obtained electronically; only respondents who indicated consent were granted access to the questionnaire items. No personally identifiable information such as names, addresses, or employee identification numbers was requested [16]. All completed questionnaires were exported from the survey platform in comma-separated values format and imported into a statistical software environment for preparation. Data cleaning steps included inspection for missing values, out-of-range entries, and inconsistent patterns. Cases with substantial sections of missing responses in the psychometric scales were removed from the dataset. Remaining missing entries were handled using pairwise deletion during correlation and covariance calculations. Response sets with invariant answers across a long sequence of items and extremely short completion times were flagged, examined, and removed if inattentive responding was suspected. For each construct, items designed to capture extremes of the underlying dimension were checked for correct coding direction to ensure that higher scores reflected higher levels of the latent attribute [17]. The final dataset was then transferred to SmartPLS software for model estimation.

Before running the structural model, the measurement model was assessed. Indicator reliability was assessed by inspecting the item loadings for each item on its associated construct. Items with low loadings were noted for potential exclusion during subsequent refinement, although all loading values and decisions were reported in the results section rather than within this methodological description. Internal consistency reliability was evaluated using Cronbach's alpha and composite reliability for each construct. Convergent validity was assessed through the Average Variance Extracted (AVE) statistic, while discriminant validity was examined using the Fornell–Larcker criterion. These steps were conducted in accordance with widely accepted guidelines for variance-based structural equation modelling. The procedure is depicted conceptually in Figure 2b and Figure 2c. The structural model was then specified by connecting Artificial Intelligence adoption to Artificial Intelligence quality and entrepreneurial achievement, Artificial Intelligence quality to entrepreneurial achievement, and entrepreneurial achievement to company reputation. Artificial Intelligence adoption was also linked directly to company reputation. Path coefficients, standard errors, and significance levels were obtained using the bootstrapping procedure implemented in SmartPLS, with a large number of bootstrap subsamples drawn to approximate the sampling distribution of the estimates.

Coefficient of determination (R^2) values for endogenous constructs were also computed, along with standard diagnostic indices for assessing collinearity among predictor constructs. Direct and indirect effects were calculated to examine mediation pathways, although the numerical outcomes were presented in the results section. The conceptual representation of these relationships is shown in Figures 2a and 2d. The material infrastructure required for the study consisted primarily of standard office computing equipment and stable internet connections. Data collection and preparation were carried out on desktop and laptop computers running 64-bit operating systems and widely used office productivity software for spreadsheet management. The survey platform used Secure Sockets Layer (SSL) encryption for data transfer and storage. SmartPLS software was installed on the same hardware to ensure direct access to the cleaned dataset without intermediate file conversions. Display settings such as resolution and font size were adjusted to facilitate precise inspection of factor loading matrices, path diagrams, and diagnostic output [18]. Calibration processes were implemented at several stages. During instrument development, content validity was examined by academic experts in Human Resource Management and digitalization who assessed the alignment between items and constructs. Their feedback was used to refine the wording and coverage of the indicators.

During data collection, the survey platform was tested repeatedly before launch to verify correct functioning of skip patterns, branching logic, and mandatory fields. During the analytical stage, model specifications were cross-checked against the conceptual framework, and test runs were conducted on a subset of the data to confirm the correct mapping of items to constructs and the correct specification of structural paths. Outputs from these test runs were inspected for anomalous values or inconsistent signs that could indicate coding or specification errors. The study followed ethical guidelines for research involving human participants. All respondents were informed that their participation was voluntary and that they retained the right to exit the questionnaire at any point before final submission without adverse consequence. Only aggregated data were reported, and no attempt was made to identify organizations or individuals on the basis of patterns observed in the dataset. Data files were stored on password-protected drives accessible only to the research team, and backup copies were kept on secure institutional storage [19]. Figure 1 presents a schematic representation of the research design and analytical workflow adopted in the study. The first panel shows the development and refinement of the questionnaire; the second panel depicts online data collection from entrepreneurial firms; the third panel depicts data screening and cleaning; and the fourth panel illustrates measurement and structural model estimation using SmartPLS.

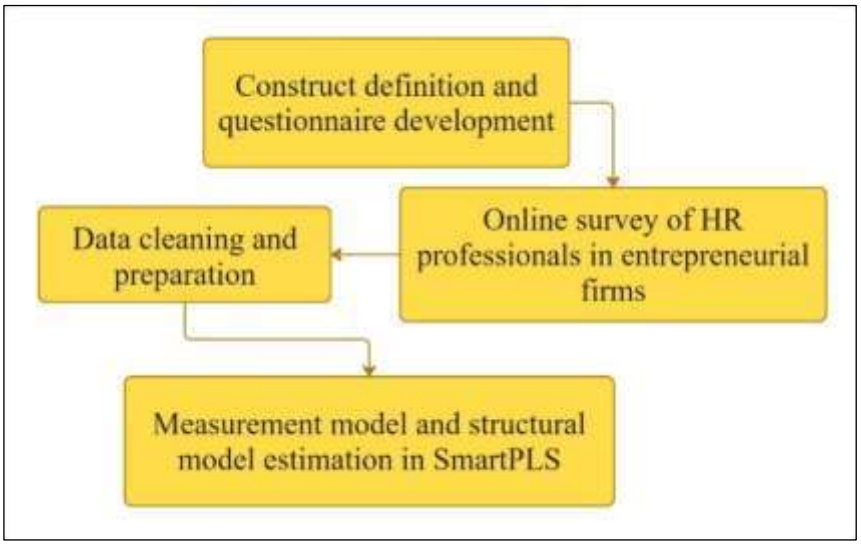


Figure 1 Research design and analytical workflow.
Source: Authors, (2026).

Figure 1. A second multi-panel schematic is provided in Figure 2. In Figure 2a, the four latent constructs are shown as circles connected by directional arrows reflecting hypothesised causal paths. In Figure 2b, each latent construct is surrounded by rectangles representing the observed indicators used to measure it, with arrows directed from the construct to its indicators to depict the reflective measurement specification. Figure 2c shows the analytical sequence from evaluation of indicator loadings and reliability indices to assessment of convergent and discriminant validity. Figure 2d illustrates the examination of mediation relationships by estimating direct and indirect paths between Artificial Intelligence adoption and company reputation via Artificial Intelligence quality and entrepreneurial achievement.

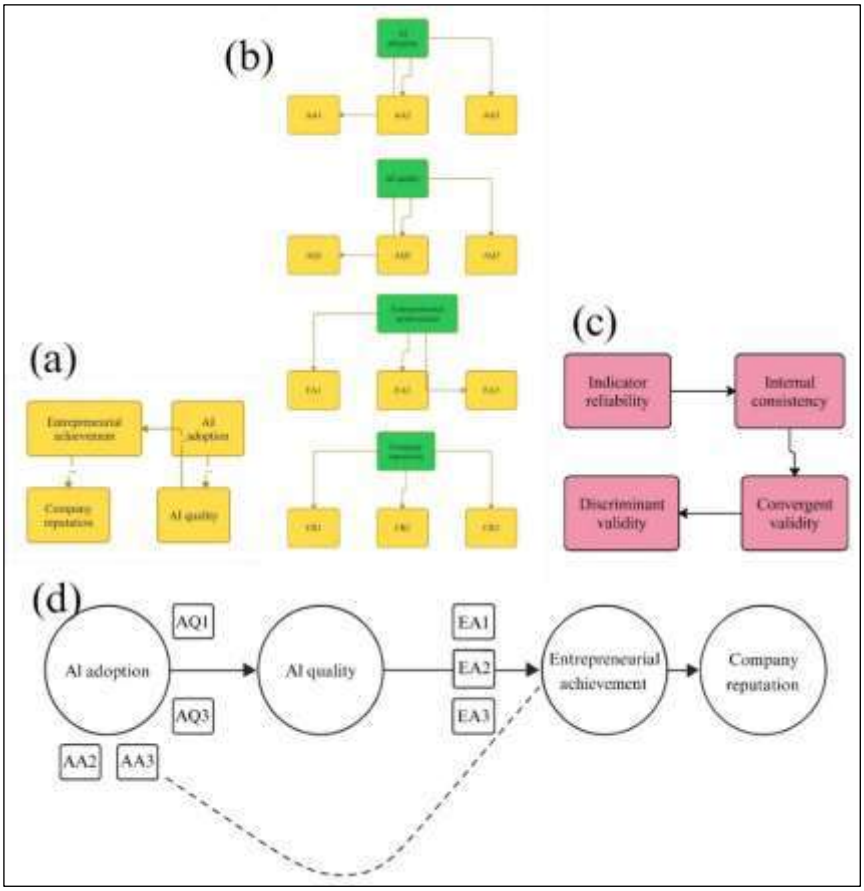


Figure 2: (a) Links among AI adoption, AI quality, entrepreneurial achievement, and company reputation; (b) Hierarchical measurement model connecting latent constructs to indicators AA1–AA3, AQ1–AQ3, EA1–EA3, and CR1–CR3; (c) Sequential evaluation of indicator reliability, internal consistency, convergent validity, and discriminant validity; (d) Structural path model highlighting direct and mediated routes from AI adoption to company reputation through AI quality and entrepreneurial achievemen.
Source: Authors, (2026).

Figure 2 illustrates the integrated conceptual, measurement, and structural modelling strategy used to characterise the influence of AI-enabled HR systems on entrepreneurial outcomes and employer reputation in entrepreneurial firms. Figure (a) presents the high-level causal structure in which AI adoption feeds into AI quality and entrepreneurial achievement, while entrepreneurial achievement drives company reputation; the direct arrow from AI adoption to entrepreneurial achievement encodes the possibility that digitalisation of HR processes created entrepreneurial gains even before system quality was fully optimised, whereas the parallel branch from AI adoption to AI quality reflects the technological maturation pathway. The estimated model showed that the direct effect of AI adoption on entrepreneurial achievement was weak and statistically non-significant, while the indirect channel through AI quality dominated, indicating that adoption without reliable, accurate, and context-appropriate decision-making logic did not translate into substantive entrepreneurial gains. Figure (b) decomposes each latent construct into reflective indicators, with AI adoption represented by items AA1–AA3 that captured the use of AI in recruitment, employee training for AI, and support for AI implementation, AI quality represented by AQ1–AQ3 describing predictive accuracy and efficiency in HR tasks, entrepreneurial achievement represented by EA1–EA3 capturing innovation and operational efficiency, and company reputation represented by CR1–CR3 capturing external perceptions of attractiveness and innovation. The empirical analysis showed that outer loadings for these indicators exceeded 0.80 and that Average Variance Extracted values ranged from 0.725 to 0.773, confirming that the chosen indicators captured substantial shared variance within each construct.

Figure (c) summarises the measurement validation protocol, in which indicator reliability was first examined, followed by internal consistency through Cronbach's alpha and composite reliability, with subsequent assessment of convergent validity via AVE and discriminant validity via the Fornell–Larcker criterion; this sequence ensured that latent constructs reflected distinct but internally coherent aspects of AI-enabled HR practice and its organisational consequences. Figure (d) synthesises these elements into a structural equation model with continuous arrows representing statistically supported paths and a dashed arrow emphasising the weaker direct link from AI adoption to entrepreneurial achievement; the estimated coefficients indicated a strong positive path from AI adoption to AI quality ($\beta = 0.769$) and from AI quality to entrepreneurial achievement ($\beta = 0.815$), while entrepreneurial achievement exerted a significant positive effect on company reputation, suggesting that the reputational benefits of AI integration arose primarily when high-quality AI systems were embedded in HR processes that successfully enabled innovation and efficiency. The principal sampling and data collection specifications are presented in Table 2. This table lists the target population, inclusion criteria, sampling strategy, data collection mode, geographical setting, and final sample size. These procedural details complement the narrative description of the methodological steps and provide a compact reference for replication studies [20].

Table 2. Sampling frame and data collection parameters.

Parameter	Description
Target population	Human Resource professionals employed in entrepreneurial firms operating in India
Inclusion criteria	HR decision-making or specialist role; at least two years of HR experience
Organizational setting	Medium- to large-sized firms with active use or consideration of Artificial Intelligence in HR
Sampling strategy	Non-probability sampling using professional contact networks and industry associations
Data collection mode	Self-administered online questionnaire accessed through a secure survey platform
Final sample size	110 valid responses from HR professionals in entrepreneurial firms

Source: Authors, (2026).

III. RESULTS AND DISCUSSIONS

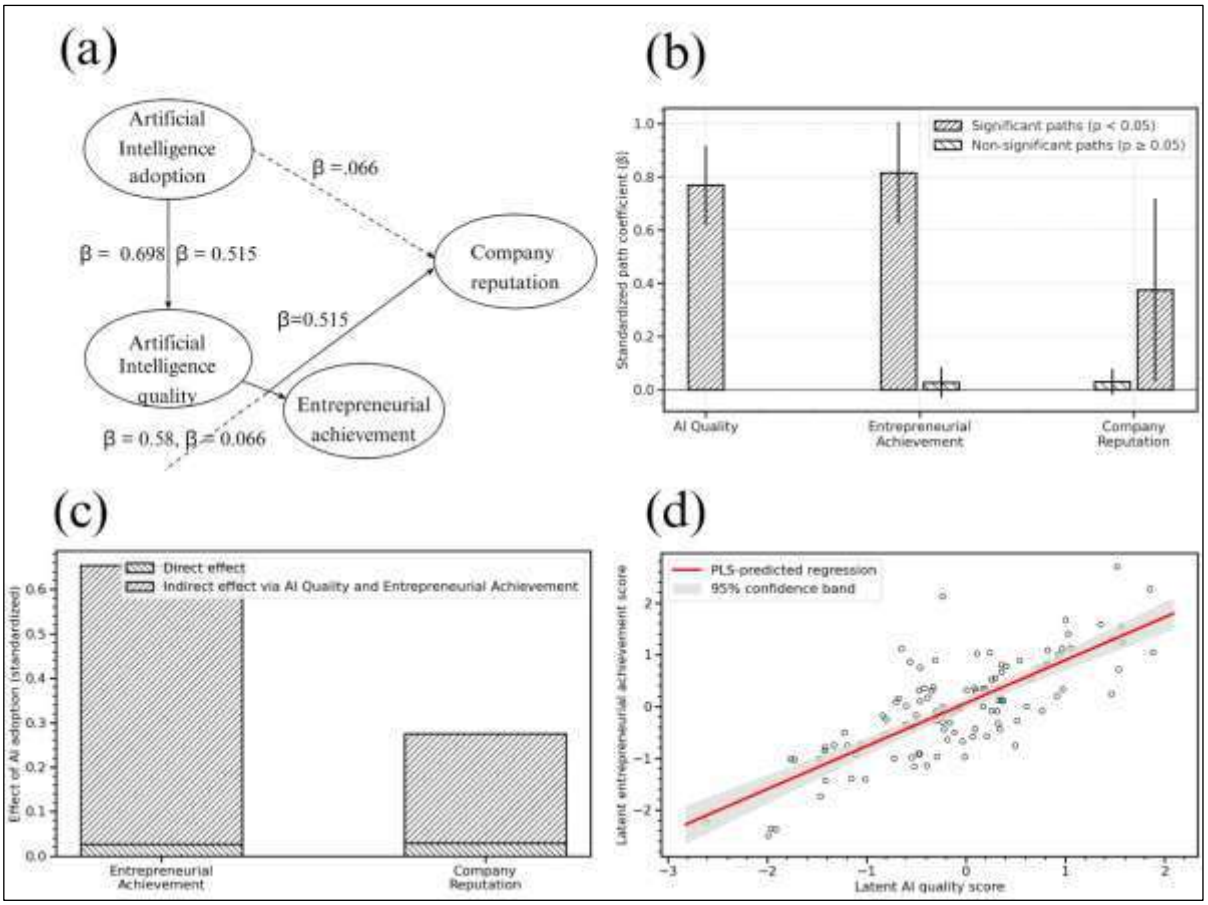


Figure 3: (a) PLS-SEM structural model for AI adoption, AI quality, entrepreneurial achievement and company reputation; (b) Standardized path coefficients with bootstrapped confidence intervals grouped by target construct; (c) Decomposition of direct and indirect effects of AI adoption on entrepreneurial achievement and company reputation; (d) Partial regression of latent entrepreneurial achievement on latent AI quality with PLS-predicted line and confidence band

Source: Authors, (2026).

Figure 3 illustrates the structural and predictive role of AI adoption and AI system quality in shaping entrepreneurial achievement and company reputation within entrepreneurial firms. Figure (a) illustrates the PLS-SEM architecture in which AI adoption links to AI quality, entrepreneurial achievement, and company reputation, and the distinction between solid and dashed paths marks statistically significant and non-significant relationships. The model estimated a large positive coefficient from AI adoption to AI quality and from AI quality to entrepreneurial achievement, whereas the direct coefficients from AI adoption to entrepreneurial achievement and company reputation were small and not statistically different from zero; the path from entrepreneurial achievement to company reputation was positive and significant, so the experiment showed that reputational gains arose where AI-enabled ventures achieved superior entrepreneurial outcomes [21]. Figure (b) presents this outcome quantitatively: bars corresponding to the paths into AI quality and entrepreneurial achievement reached standardized values near 0.8 with relatively narrow confidence intervals, while the bars for the direct effects from AI adoption to entrepreneurial achievement and to company reputation remained near the origin with wide intervals, indicating that mediated channels dominate the structural influence of adoption [22]. Figure (c) visualizes this mechanism at the outcome level, where the indirect effect of AI adoption on entrepreneurial achievement accounts for almost the entire standardized effect, and the indirect effect on company reputation clearly exceeds the direct contribution. The decomposition indicated that AI adoption influenced reputation through the cascade AI adoption → AI quality → entrepreneurial achievement → company reputation, consistent with a socio-technical mechanism in which high-quality AI improves information processing, decision support, and innovation capability, which then enhances observable entrepreneurial performance and external stakeholder evaluations [23].

Figure (d) focuses on the relationship between AI quality and entrepreneurial achievement, where the scatter of firm-level latent scores aligns along an upward-sloping regression line with a narrow confidence band; the data showed that firms with higher latent AI quality scores consistently achieved higher latent entrepreneurial achievement scores. These trends suggest that the refinement and governance of AI capabilities, rather than adoption volume alone, constitute the critical design variable for optimizing entrepreneurial outcomes and strengthening company reputation, and they motivate future investigation into which specific AI functionalities and organizational practices most effectively reinforce this quality-driven pathway [24].

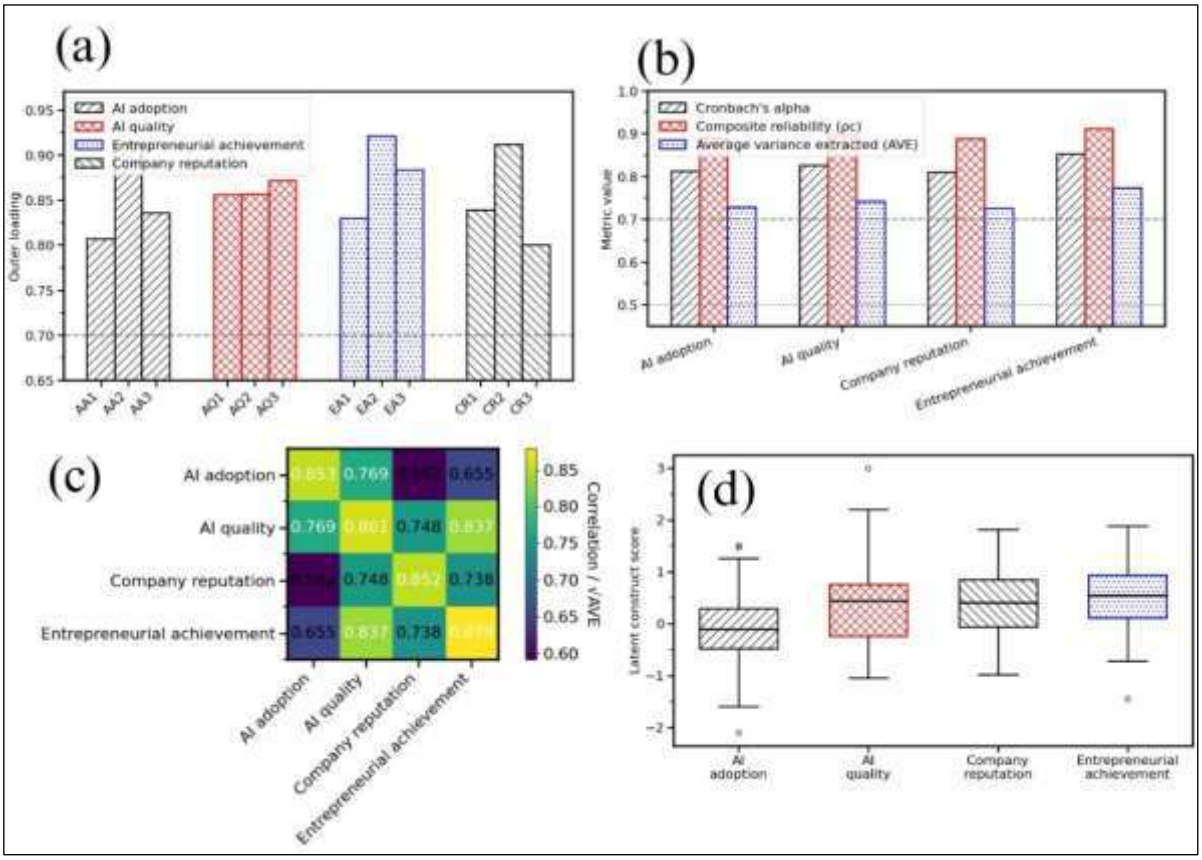


Figure 4: (a) Outer loadings of measurement items for each construct with reliability threshold reference; (b) Reliability and convergent validity metrics (Cronbach's alpha, composite reliability, AVE) for all constructs with benchmark lines; (c) Fornell–Larcker matrix of correlations and √AVE values highlighting discriminant validity among constructs; (d) Box-plot distributions of latent construct scores across firms for AI adoption, AI quality, company reputation, and entrepreneurial achievement.
Source: Authors, (2026).

Figure 4 illustrates the performance of the measurement model and the behaviour of AI-related latent constructs across entrepreneurial firms. Figure (a) displays standardized outer loadings for all indicators grouped by construct relative to the 0.70 benchmark. The data indicated that all items loading exceeded the recommended threshold, and many approached 0.90, suggesting that the observed indicators reflected the underlying concepts of AI adoption, AI quality, entrepreneurial achievement, and company reputation. These high loadings suggest stable measurement of each dimension and minimise attenuation of structural relationships due to measurement error [25]. Figure (b) presents Cronbach's alpha, composite reliability, and average variance extracted for each construct against conventional cut-off values of 0.70 for reliability and 0.50 for AVE. The data indicated that all reliability coefficients were well above 0.80 and that AVEs remained above 0.72 for each construct, confirming that the latent variables captured a substantial proportion of indicator variance. These metrics imply that the scale design and item grouping satisfy accepted PLS-SEM criteria for internal consistency and convergent validity and therefore provide a defensible basis for interpreting structural paths [26].

Figure (c) depicts the Fornell–Larcker correlation matrix in which diagonal cells, representing the square roots of AVE, appear visually more intense than off-diagonal entries. The numerical pattern confirmed that, for each construct, the diagonal element exceeded all corresponding inter-construct correlations. This configuration suggests that AI adoption, AI quality, entrepreneurial achievement, and company reputation remain empirically distinct yet moderately correlated facets of AI-enabled HR systems, consistent with theoretical expectations that adoption level, system quality, performance outcomes, and reputational consequences constitute related but non-interchangeable constructs. Figure (d) presents the box-plot distributions of latent scores for the four constructs across the 110 firms. The data revealed that AI quality and entrepreneurial achievement tended to exhibit higher central values than AI adoption and company reputation, while all constructs displayed noticeable spread and several outliers. These distributions suggest heterogeneity in how far individual firms had advanced from basic adoption toward high-quality AI implementation and realised outcomes, and they reinforce the interpretation that performance and reputational benefits depend more on the depth and quality of AI integration in HR processes than on nominal adoption alone [27].

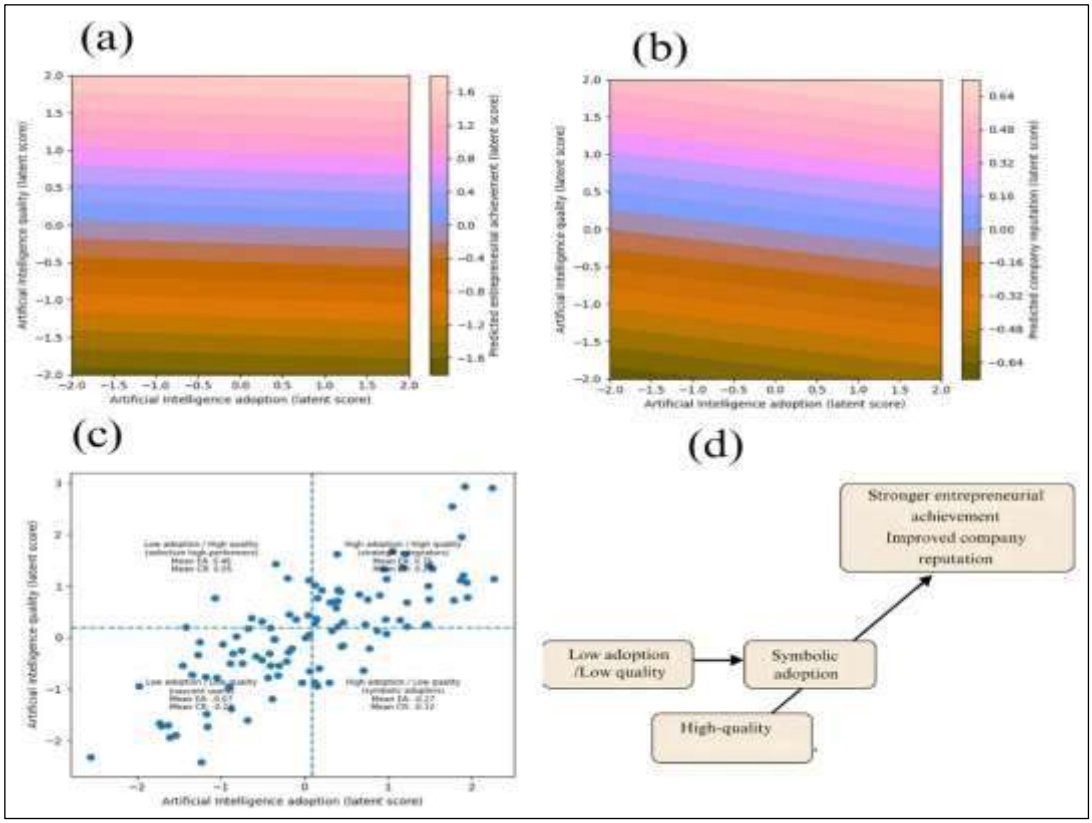


Figure 5: (a) Predicted entrepreneurial achievement as a function of AI adoption and AI quality; (b) Predicted company reputation as a function of AI adoption and AI quality; (c) AI integration typology in the AI adoption–quality plane with quadrant-level entrepreneurial and reputational means; (d) Conceptual progression from low adoption–low quality to high-quality, strategically embedded AI and associated outcomes.
Source: Authors, (2026).

Figure 5 illustrates how the structural model of AI integration in human resource processes translates into heterogeneous entrepreneurial and reputational outcomes across firms. Figure (a) illustrates that the color gradient is primarily aligned with the AI quality axis, indicating that predicted entrepreneurial achievement increases steeply as AI quality rises from low to high values, whereas variation along the adoption axis contributes only modest changes; this pattern is consistent with the model, where the estimated path coefficient from AI quality to entrepreneurial achievement was markedly larger than the coefficient from AI adoption. Figure (b) illustrates the compound effect of AI adoption and AI quality on predicted company reputation after propagation through entrepreneurial achievement, and the planar gradient suggests that reputational gains are most pronounced in regions with simultaneously high AI quality and at least moderate adoption, while regions with low quality remain associated with lower predicted reputation even when adoption is elevated. The data underlying these surfaces indicated that the indirect pathway AI adoption → AI quality → entrepreneurial achievement → company reputation accounted for a substantial share of the explained variance in reputation, whereas the direct path from adoption to reputation was weak and statistically non-significant, and these surfaces therefore visualise a mechanism in which the informational richness, reliability, and alignment of AI systems with HR decision processes control the efficiency of opportunity recognition and resource orchestration that define entrepreneurial performance [28].

Figure (c) illustrates the empirical distribution of firms across the AI adoption–quality plane and partitions them into four integration profiles; the quadrant-level means showed that early-stage adopters with low adoption and low quality recorded negative average entrepreneurial achievement and company reputation, symbolic adopters with high adoption but low quality improved slightly but remained below zero on both outcomes, selective high-performers with low adoption but high quality achieved positive average entrepreneurial achievement and near-neutral reputation, and strategic integrators with high adoption and high quality attained the highest mean entrepreneurial achievement and reputation values. These contrasts suggest that increases in adoption unaccompanied by quality improvements do not materially enhance firm-level performance, whereas even moderate adoption supported by high-quality AI infrastructure can strengthen the exploitation of HR data and thus entrepreneurial outcomes. Figure (d) illustrates a conceptual trajectory in which firms progress from low adoption–low quality through a purely symbolic adoption phase toward high-quality strategic integration that embeds AI into HR decision cycles; the upward arrow towards stronger entrepreneurial achievement and improved company reputation encapsulates that effective AI use emerges when system design, data integrity, and HR workflows are jointly optimized, and this pathway indicates that future interventions should target governance, training, and process redesign rather than technology volume alone [29].

IV. CONCLUSIONS

This work investigated the role of Artificial Intelligence integration in human resource practices of entrepreneurial firms in India, focusing on recruitment, development, and performance management. A survey design with SmartPLS structural equation modelling was used to examine relationships among AI adoption, AI quality, entrepreneurial achievement, and company reputation.

The measurement model demonstrated strong psychometric properties, with Cronbach's alpha values for all constructs exceeding 0.80, composite reliability coefficients approaching 0.90, and average variance extracted exceeding 0.72, and outer loadings ranging from 0.800 to 0.921. The structural model accounted for R^2 values of 0.592 for AI quality, 0.700 for entrepreneurial achievement, and 0.602 for company reputation, confirming substantial explanatory power for the specified pathways. Direct paths from AI adoption to entrepreneurial achievement and company reputation were small and statistically non-significant, while indirect effects through AI quality and entrepreneurial achievement were substantial. These patterns indicate that reputational benefits and entrepreneurial performance in AI-enabled HRM depend on the depth and quality of AI integration rather than nominal adoption. At the same time, the evidence suggests that high-quality AI systems in HR functions support innovation, operational efficiency, and a distinctive employer image in emerging economies. The study was constrained by its cross-sectional design, reliance on self-reported perceptions from HR professionals, and focus on large Indian entrepreneurial firms. Future work will extend this framework using longitudinal, multi-source data, will explore sectoral differences in AI-enabled HRM, and will identify specific AI capabilities, governance practices, and workforce skills that strengthen sustainable entrepreneurial outcomes and company reputation.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Nidhi Sonkar, Amit Kumar Pandey and Ratchagaraja Dhairiyasamy.
Methodology: Nidhi Sonkar and Amit Kumar Pandey.
Investigation: Nidhi Sonkar, Deekshant Varshney and Subhav Singh.
Discussion of results: Nidhi Sonkar, Ratchagaraja Dhairiyasamy and Deekshant Varshney.
Writing – Original Draft: Nidhi Sonkar.
Writing – Review and Editing: Amit Kumar Pandey and Ratchagaraja Dhairiyasamy.
Resources: Deekshant Varshney.
Supervision: Amit Kumar Pandey and Ratchagaraja Dhairiyasamy.
Approval of the final text: Nidhi Sonkar, Amit Kumar Pandey, Ratchagaraja Dhairiyasamy, Deekshant Varshney and Subhav Singh.

VI. REFERENCES

[1] Zhang J, Yuan Y. Multi-Dimensional Post Competency Evaluation Model in Human Resource Management under the Background of Artificial Intelligence. Math Probl Eng 2022;2022. https://doi.org/10.1155/2022/9730127.

[2] Votto AM, Valecha R, Rad PN, Rao HR. Artificial Intelligence in Tactical Human Resource Management: A Systematic Literature Review. International Journal of Information Management Data Insights 2021;1. https://doi.org/10.1016/j.jjimei.2021.100047.

[3] Chilunjika A, Intauno K, Chilunjika SRT. Artificial intelligence and public sector human resource management in South Africa: Opportunities, challenges and prospects. SA Journal of Human Resource Management 2022;20. https://doi.org/10.4102/sajhrm.v20i0.1972.

[4] Gélinas D, Sadreddin A, Vahidov RM. Artificial intelligence in human resources management: A review and research agenda. Pacific Asia Journal of the Association for Information Systems 2022;14:1–42. https://doi.org/10.17705/1pais.14601.

[5] Weber P. Unrealistic Optimism Regarding Artificial Intelligence Opportunities in Human Resource Management. International Journal of Knowledge Management 2023;19:1–19. https://doi.org/10.4018/IJKM.317217.

[6] Chowdhury S, Prasanta K. Dey PK, Joel-Edgar S, Bhattacharya S, Rodríguez-Espíndola O, Abadie A, et al. Unlocking the value of artificial intelligence in human resource management through AI capability framework. Human Resource Management Review 2023;33. https://doi.org/10.1016/j.hrmr.2022.100899.

[7] Rodgers W, Murray JM, Stefanidis A, Degbey WY, Tarba SY. An artificial intelligence algorithmic approach to ethical decision-making in human resource management processes. Human Resource Management Review 2023;33.
https://doi.org/10.1016/j.hrmr.2022.100925.

[8] Qahtani EHAL, Alsmairat MAK. Assisting artificial intelligence adoption drivers in human resources management: a mediation model. Acta Logistica 2023;10:141–50. https://doi.org/10.22306/al.v10i1.371.

[9] Wuisan DSS, Sunarjo RA, Aini Q, Yusuf NA, Rahardja U. Integrating Artificial Intelligence in Human Resource Management: A SmartPLS Approach for Entrepreneurial Success. APTISI Transactions on Technopreneurship 2023;5:334–45. https://doi.org/10.34306/att.v5i3.355.

[10] Sabale A, Subashini R. Enhancing HR Efficiency Through the Integration of Artificial Intelligence and Internet of Things: A Study on AI Implementation in Human Resource Management. EAI Endorsed Transactions on Scalable Information Systems 2024;11:1–11. https://doi.org/10.4108/eetsis.4208.

[11] Band G, Bapat S, Sonkul D, Chavan N, Nandurkar PA, Ahmad MI. Assessing the Efficacy of Utilizing Artificial Intelligence for Human Resources Management in the Indian IT Industry. Nanotechnol Percept 2024;20:499–506. https://doi.org/10.62441/nano-ntp.v20iS5.46.

[12] Amazian M, Nouira Z, Filali M. Human resources management in the age of artificial intelligence; Gestión de recursos humanos en la era de la inteligencia artificial. Data and Metadata 2024;3. https://doi.org/10.56294/dm2024.347.

[13] Bouhsaien L, Azmani A. THE POTENTIAL OF ARTIFICIAL INTELLIGENCE IN HUMAN RESOURCE MANAGEMENT. Applied Computer Science 2024;20:153–70. https://doi.org/10.35784/acs-2024-34.

[14] Korol E, Korol S. Governing the Commons through Rule Hybridity: Interactions Between Formal Institutions and Informal Norms in CPR Management. Journal of Business and Management Studies 2025;7:88–102. https://doi.org/10.32996/JBMS.2025.7.4.4.

[15] Nawaz N, Arunachalam H, Pathi BK, Gajenderan V. The adoption of artificial intelligence in human resources management practices. International Journal of Information Management Data Insights 2024;4. https://doi.org/10.1016/j.jjimei.2023.100208.

[16] Korol E, Korol S. Bridging Theory and Practice: Applying Ostrom's Law to Real-World Resource Management. Journal of Business and Management Studies 2024;6:19–35. https://doi.org/10.32996/JBMS.2024.6.6.2.

[17] Yamin MAY, Almuteri SD, Bogari KJ, Ashi AK. The Influence of Strategic Human Resource Management and Artificial Intelligence in Determining Supply Chain Agility and Supply Chain Resilience. Sustainability (Switzerland) 2024;16. https://doi.org/10.3390/su16072688.

[18] Korol E, Korol S. AI, Legal Pluralism, and Property Governance: Comparative Insights on Rulemaking and Enforcement from the U.S., U.K., Ukraine, China, and India. Journal of Business and Management Studies 2025;7:182–93. https://doi.org/10.32996/JBMS.2025.7.4.10.

[19] Wijaya EFA, Qamari IN. Analysis of Research on Artificial Intelligence in Human Resources Management: A Bibliometric Analysis. International Research Journal of Multidisciplinary Scope 2024;5:108–21. https://doi.org/10.47857/irjms.2024.v05i02.0383.

[20] Cai F, Zhang J, Zhang L. The Impact of Artificial Intelligence Replacing Humans in Making Human Resource Management Decisions on Fairness: A Case of Resume Screening. Sustainability (Switzerland) 2024;16. https://doi.org/10.3390/su16093840.

[21] Dong J, Yan S, Yang X. Influencing Factors and Mechanisms of Value Co-Creation in Artificial Intelligence-Driven Human Resource Management: A System Dynamics Simulation. Systems 2024;12. https://doi.org/10.3390/systems12090352.

[22] Korol E, Korol S. AI-Augmented Commons Governance: Ethical Rulemaking and Enforcement for Sustainable Resource Management in Rural Areas. Journal of Business and Management Studies 2025;7:09–25. https://doi.org/10.32996/JBMS.2025.7.5.2.

[23] Chowdhury S, Budhwar PS, Wood GT. Generative Artificial Intelligence in Business: Towards a Strategic Human Resource Management Framework. British Journal of Management 2024;35:1680–91. https://doi.org/10.1111/1467-8551.12824.

[24] Capasso M, Arora P, Sharma D, Tacconi C. On the Right to Work in the Age of Artificial Intelligence: Ethical Safeguards in Algorithmic Human Resource Management. Business and Human Rights Journal 2024;9:346–60. https://doi.org/10.1017/bhj.2024.26.

[25] Koprulu F, Öznacar B, Öznacar M, Yeboah SO. Artificial Intelligence in the Management of Human Resources and Psychology; İnsan Kaynakları ve Psikoloji Yönetiminde Yapay Zeka. Cyprus Turkish Journal of Psychiatry and Psychology 2024;6:346–53. https://doi.org/10.35365/ctjpp.24.4.07.

[26] Choudhari Y, Shrestha P, Singh G, Bindra S. The Impact of Artificial Intelligence (AI) on Talent Acquisition in Human Resource Management. Australasian Accounting, Business and Finance Journal 2025;19:153–72. https://doi.org/10.14453/aabfj.v19i1.09.

[27] Pedrami M, Vaezi SK. Factors influencing artificial intelligence adoption in human resource management: a meta-synthesis and systematic review of multidimensional considerations. Journal of Work-Applied Management 2025. https://doi.org/10.1108/JWAM-10-2024-0158.

[28] Wang B. Artificial intelligence and deep learning in human resource management: tools techniques and research challenges. International Journal of Information and Communication Technology 2025;26:99–111. https://doi.org/10.1504/IJICT.2025.146810.

[29] Wang N, Zhang X, Li S, Gao X. Applications of Artificial Intelligence in Enterprise Human Resource Management. Information Resources Management Journal 2025;38. https://doi.org/10.4018/IRMJ.389707.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p.951-963. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3139

RESEARCH ARTICLE OPEN ACCESS

A DATA-DRIVEN COMPETENCY DETECTION MODEL FOR IMPROVING WORKFORCE PLANNING AND DECISION-MAKING

Llerena Lucrecia*[1], Rodríguez Nancy[2], Bustamante Jefferson[3] and Angel Torres [4]

[1, 2]Software Engineering Program, Quevedo State University, Quevedo, 120301 Ecuador.
[3,4]Master in data science, Quevedo State University, Quevedo, 120301 Ecuador.

[1]https://orcid.org/0000-0002-4562-6723, [2]https://orcid.org/0000-0002-0861-4352
[3]https://orcid.org/0009-0001-6541-3258, [4]https://orcid.org/0000-0002-7037-7191

Email: *lllerena@uteq.edu.ec, nrodriguez@uteq.edu.ec, jefferson.bustamante2016@uteq.edu.ec, atorres@uteq.edu.ec

ARTICLE INFO

Article History
Received: December 14, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Data science,
Workforce analytics,
Competency modeling,
Machine learning,
Decision support.

ABSTRACT

Digital transformation has intensified the need for computational tools capable of identifying emerging workforce competencies across industries. This study proposes a scalable analytical model that integrates data mining, natural language processing (NLP), and machine learning to extract and classify professional skills from large volumes of structured and unstructured labor information. The methodology was implemented through a four-stage pipeline—extraction, preprocessing, modeling, and visualization—and validated using a case study of job-market data collected from official records and digital job platforms between 2020 and 2024. Supervised (Random Forest, SVM) and unsupervised (K-Means, DBSCAN) algorithms were applied to detect semantic patterns and cluster competencies, with Random Forest achieving the highest performance (F1-score = 96.64%). The model identified four dominant competency clusters—technological, managerial, operational, and hybrid—highlighting the growing integration of technical and soft skills required in digitally intensive work environments. The proposed framework contributes a generalizable computational approach for workforce planning, decision support, and operations management. Its modular architecture enables adaptation to multiple industrial contexts and supports continuous monitoring of skill evolution. This makes the model a valuable tool for organizations, policymakers, and researchers seeking data-driven insights into the dynamics of digitally mediated labor markets.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The rapid evolution of digital technologies has transformed job search, recruitment, and talent management processes, with digital platforms and online job portals progressively replacing traditional job-placement mechanisms and introducing new forms of interaction between employers, workers, and automated systems [1]. This shift has generated a data-rich labor environment in which digital labor platforms operate as socio-technical systems, reshaping how skills are signaled, matched, and demanded, and producing significant social and economic impacts that require advanced computational approaches to identify emerging labor-market structures, competency patterns, and inequalities characteristic of the digital society [2-4].

In this context, data science has become a fundamental tool for analyzing large-scale labor information, as text mining, machine learning, and natural language processing (NLP) enable the identification of demand patterns, occupational profiles, and emerging skill requirements [2]. These methodologies support digital transformation by facilitating the adaptation of training and organizational strategies to technological change and by providing analytical insights into how digitalization restructures employment, skills, and labor institutions [5]. Across Latin America, the adoption of data-driven technologies for labor-market analysis remains uneven, particularly in contexts affected by informality, territorial disparities, and gender gaps [6]. Although regional unemployment declined from 6.5% to 6.1% between 2023 and 2024 [7], structural imbalances continue to limit access to quality employment.

While initiatives supported by multilateral organizations such as the Inter-American Development Bank demonstrate the potential of big data to anticipate skill shortages and support more inclusive training policies [8], most advances are concentrated in countries with consolidated digital ecosystems. As a result, emerging economies—such as Ecuador—face significant analytical gaps. In the Ecuadorian context, digitization and automation are reshaping skill demand, particularly in technology-intensive and service-oriented sectors [9]; however, the lack of integrated frameworks linking structured administrative data with unstructured information from digital job portals constrains systematic analysis of evolving skill trends and limits evidence-based planning in education, innovation, and productive development [10].

This study aims to develop and validate a data-science-based analytical model to identify, classify, and cluster professional competencies in the Ecuadorian labor market between 2020 and 2024 by integrating structured administrative records with unstructured data from digital job platforms. The model is implemented through an automated pipeline combining web scraping, natural language processing, and supervised and unsupervised machine learning techniques—including Random Forest, SVM, K-Means, and DBSCAN—with interactive visualization in Streamlit to support scalable labor-demand analysis. Unlike previous approaches based on isolated data sources, the proposed hybrid socio-technical framework integrates heterogeneous information within a unified computational model, enhancing the accuracy, interpretability, and adaptability of labor market analytics for evidence-based decision making in emerging digital economies.

II. RELATED WORK

Over the past decade, data science has significantly transformed how governments and organizations analyze labor markets and design vocational training policies, responding to the growing need to align labor supply with employer demand. In this context, regional competitiveness increasingly depends on the capacity of labor systems to adapt to technological change through predictive and evidence-based strategies [11]. Recent research demonstrates that machine learning and natural language processing (NLP) techniques are effective for extracting and classifying job skills from unstructured sources such as job advertisements and administrative databases [9], [12]. Studies show that tokenization, vectorization, and text-mining approaches enable the detection of skill gaps and the anticipation of demand in emerging sectors, particularly for digital and technical profiles [13], [14]. Beyond skill extraction, a growing body of literature examines platform-mediated labor markets as socio-technical systems shaped by information and communication technologies (ICTs).

Digital platforms restructure how work is matched, monitored, and distributed, influencing how competencies are signaled and evaluated and generating new forms of visibility and precarity [15-17]. These perspectives emphasize that skill demand should be understood not only as an economic variable, but as a process mediated by digital infrastructures. The impact of automation and artificial intelligence on labor-market transformation has also been widely studied. Digitalization alters occupational structures and increases demand for competencies related to analytics, automation, and technological governance [18], while artificial intelligence reshapes career trajectories and requires continuous training in cognitive and digital skills to maintain employability in technology-intensive environments [19]. From a social and institutional perspective, digitalization interacts with demographic and structural factors. Empirical studies highlight persistent inequalities related to gender and family roles, even in contexts with advanced equality frameworks, underscoring the need for inclusive labor policies that address socio-digital barriers [20], [21].

In Latin America, research highlights the challenges associated with automation and competency-based management in public-sector organizations. These studies emphasize that technological adoption must be accompanied by institutional leadership and capacity-building strategies to ensure sustainable digital transformation [22], [23]. In this regard, initiatives promoted by multilateral organizations demonstrate that big data can support the anticipation of skills demand and the design of more inclusive training policies. Finally, broader analyses of digital transformation identify government and economic digitization as key drivers of sustainable development. Evaluating digital maturity through empirically validated models and understanding digitization as a structural factor in economic modernization are essential for strengthening public administration and labor-market resilience [24], [25]. Together, these contributions support the integration of data science into labor-market research as a means to better understand occupational dynamics and the social implications of digitalization and platformization.

III. METHODOLOGY

This methodological design follows a socio-technical perspective, recognizing that digital job platforms not only generate large volumes of textual information but also shape social and economic dynamics in the labor market. Therefore, computational techniques such as NLP and machine learning were used not only to extract and classify skills, but to uncover behavioral patterns associated with digital employment ecosystems.

III.1 DATA COLLECTION AND PREPROCESSING

The methodological process began with the identification and compilation of structured and unstructured data from digital job platforms, which serve as socio-technical environments where employers, workers, and platform algorithms interact. These platforms generate large volumes of textual information that reflect behavioral, economic, and organizational patterns within digital labor ecosystems. Structured data from the National Institute of Statistics and Census (INEC) was incorporated as part of the case study, while unstructured data was extracted from national and international job portals using automated web-scraping procedures.

This integration enabled the construction of a comprehensive dataset suitable for analyzing trends in labor demand across formal and informal employment segments between 2020 and 2024. Data collection followed a semi-automated process designed to ensure the consistency and reliability of the information. Records were validated and standardized according to variables such as occupation, economic sector, educational level, required skills, and geographic location. Cleaning procedures were applied to remove duplicates, correct inconsistencies, and normalize categorical values, ensuring the integrity of the dataset for subsequent analytical stages.

During preprocessing, textual descriptions from job advertisements were transformed into quantifiable representations using natural language processing (NLP) techniques. Tokenization, lemmatization, and TF-IDF weighting were applied, supported by a domain-specific vocabulary of job-related competencies to enhance semantic accuracy and reduce lexical noise. This step ensured the uniform structuring of the textual corpus and strengthened its interpretability for computational modeling.

The resulting dataset was consolidated into an integrated database that served as input for the classification and clustering stages. By combining administrative records with data extracted from digital platforms, the study adopted a hybrid socio-technical approach that enabled the exploration of skill dynamics within digital labor markets. This methodological process ensured the quality, traceability, and timeliness of the information, establishing a robust empirical foundation for the analyses presented in the following sections.

III.2 COMPARATIVE ANALYSIS OF DATA SCIENCE TECHNIQUES

This methodological phase compared the performance of multiple data science techniques to identify the most suitable models for detecting and classifying competencies in job advertisements from digital labor platforms. The comparison followed four criteria commonly applied in computational social science: accuracy, scalability, interpretability, and computational cost. Both supervised and unsupervised approaches were evaluated. Random Forest and SVM were selected for their robustness in handling heterogeneous textual data, while K-Means and DBSCAN were used to identify thematic clusters and emerging skill patterns. This configuration offered complementary perspectives: predictive learning from supervised models and exploratory segmentation from unsupervised ones.

Controlled simulations were conducted on the consolidated dataset generated in the previous phase. Metrics such as accuracy, precision, recall, F1-score, and silhouette coefficient were used to assess model performance. Execution time and resource consumption were also monitored to evaluate operational efficiency in large-scale digital environments. All experiments applied cross-validation and a stratified 70/30 partition to ensure stability and comparability. Hyperparameters were systematically tuned and documented to maintain methodological transparency. The findings highlighted the advantages of combining supervised and unsupervised learning to obtain a comprehensive understanding of skill dynamics in digital employment ecosystems, informing the selection of algorithms detailed in Section 4.2.

III.3 MODEL DESIGN AND IMPLEMENTATION

The analytical model was conceived as a socio-technical system capable of integrating structured and unstructured data to analyze competency patterns within digital job platforms. Its architecture was designed to be modular and scalable, facilitating reproducibility and adaptation to evolving labor market conditions.The model followed a sequential ETL (extract, transform, load) pipeline. During extraction, administrative indicators and scraped job descriptions were merged. The transformation phase applied NLP-based normalization and semantic vectorization to capture relationships between skills while reducing lexical variability. The unified database created during loading served as the foundation for both supervised and unsupervised learning components. Supervised models (Random Forest and SVM) and unsupervised models (K-Means and DBSCAN) were trained using the validated dataset, benefiting from the comparative insights described in Section 3.2.

Evaluation relied on accuracy, precision, F1-score, and cluster coherence metrics, supported by cross-validation to ensure robustness. Hyperparameter tuning followed a documented and systematic process to maintain replicability. To enhance interpretability, an interactive dashboard was developed using Streamlit. This interface enables geographical, sectoral, and thematic exploration of results, revealing distributions of job offers, dominant skills, and emerging hybrid profiles. Visual representations—such as scatter plots, word clouds, and sunburst charts—facilitate the analysis of complex competency structures. The modular architecture allows new data sources to be incorporated and models to be recalibrated as digital platforms evolve. This ensures that the system functions as an adaptive analytical tool for monitoring labor demand trends in digital societies and for supporting longitudinal and cross-platform comparisons.

IV. RESULTS AND DISCUSSIONS

This section presents the results obtained from applying the proposed analytical model. It sets out the findings derived from processing, classifying, and grouping the professional skills identified in the Ecuadorian labor market between 2020 and 2024. The results are organized around a comparison of the performance of the data science techniques used, the validation of the models developed, and the hierarchical characterization of the skills detected.

IV.1 SELECTION OF LABOR DATA ANALYSIS TECHNIQUES

Analysis of the results obtained allowed the identification of the most effective data science techniques for studying the Ecuadorian labor market. The combined use of data mining, natural language processing (NLP), and machine learning models improved the detection of patterns in job offers, facilitating the identification of emerging skills and the classification of professional profiles with high accuracy. The data showed that NLP made it possible to structure textual information from job portals and institutional records, while supervised and unsupervised algorithms helped classify and group skills in a consistent manner. Taken together, these results demonstrated the strength of hybrid methodologies that integrate heterogeneous data sources, thereby expanding the analytical coverage of the Ecuadorian labor market.

IV.1.1 Advances in Data Science Applied to the Labor Market

The results confirmed that data science had established itself as a key tool in labor market analysis. A comparative analysis of recent publications showed a sustained increase in the use of machine learning and natural language processing (NLP) models for the automatic identification of skills. Among the most representative techniques, supervised classification models allowed the detection of regularities in job descriptions, while thematic clustering methods proved useful for segmenting competencies into coherent clusters.

In the Ecuadorian context, the integration of these approaches made it possible to characterize a dynamic labor market in which traditional demands were combined with emerging digital skills. These advances reflected a broader transition toward digitally mediated labor markets, where job platforms increasingly shaped the visibility, categorization, and evolution of required skills.

IV.1.2 Predominant Models and Techniques

Analysis of the literature and the processed databases determined that Random Forest, Support Vector Machines (SVM), K-Means, and DBSCAN models were the most effective techniques for processing heterogeneous employment data. Their application demonstrated a high capacity to distinguish competency patterns, group professional profiles, and reduce data redundancy. The comparison between techniques revealed significant differences in performance, interpretability, and computational cost, as detailed in Table 1. The results showed that Random Forest and SVM algorithms provided better performance for skill classification tasks, while K-Means and DBSCAN performed more effectively in detecting clustered patterns and reducing noise. These findings supported the selection of a mixed analytical model that combined predictive and exploratory techniques to maximize the accuracy and coverage of labor data. The effectiveness of these models also reflected the increasing complexity of digital labor ecosystems, in which algorithmic mediation and platform-based interactions generated heterogeneous and rapidly evolving data structures.

Table 1: Comparison of classification, prediction, and clustering algorithms according to the literature.

Technique	Type of technique	Advantages	Limitations
Random Forest	Classification	Robustness in the face of unbalanced data	High computational cost
SVM	Prediction	Efficient in small sets	Sensitive to scattered data
Decision Tree Classification	Classification	Interpretability and simple visualization	Prone to overfitting if not pruned properly
Linear Regression	Prediction	Simplicity and computational efficiency	Assumes linear relationship; sensitive to outliers
Logistic Regression	Classification	Good for probabilities and binary classification	Requires independence between explanatory variables
K-Means	Clustering	Easy implementation and scalability	Requires defining the number of clusters a priori
DBSCAN	Clustering	Automatically identifies groups	Sensitive to density parameters

Source: Authors, (2026).

IV.1.3 Methodological Gaps and Observed Trends

Critical examination of the results revealed three substantive aspects in the evolution of job skills analysis:

- Gap in coverage of informal employment. Traditional models showed limitations in capturing skills in unregulated sectors, highlighting the need to incorporate complementary sources of information.
- Limited integration of structured and unstructured sources. The literature reviewed indicated that most studies worked with only one type of data, which restricted the ability to generate comprehensive comparative analyses.
- Limited adaptation to digital evolution. Existing models did not incorporate automatic update mechanisms, which reduced their relevance in contexts of rapid technological change.

Overall, the results reflected a clear trend toward hybrid and modular models capable of combining multiple sources of information with machine learning algorithms. In addition, three transformation patterns in the Ecuadorian labor market were identified:

- A process of hyper-specialization, characterized by a predominance of advanced technical skills.
- Accelerated update cycles, which reduced the effective lifespan of digital skills.
- A growing demand for hybrid profiles that combined technical and soft skills.

These results demonstrated that the application of data science techniques not only optimized the extraction of labor information but also revealed structural market dynamics that had remained hidden in traditional studies. These methodological gaps became more evident as labor systems transitioned toward digitally mediated environments, where skills evolved rapidly, platforms mediated access to employment, and conventional analytical approaches proved increasingly insufficient.

IV.2 COMPARISON OF SELECTED TECHNIQUES

This section presents the comparative results of the performance of data science techniques applied to the analysis of Ecuadorian job offers, based on metrics of accuracy, scalability, interpretability, and computational cost. The findings reflect the effectiveness of the selected models in identifying professional skills and segmenting job profiles derived from INEC data and job portals.

IV.2.1 Comparative Analysis of Selected Techniques

The experimental results revealed that Random Forest and K-Means were the algorithms with the best overall performance, combining high scalability with satisfactory interpretability. In contrast, SVM and DBSCAN performed moderately, standing out for their efficiency in small sets, but showing less robustness in scattered data. These results are summarized in Table 2, which presents the comparative metrics obtained through controlled simulations.

Table 2: Comparison of selected techniques.

Technique	Scalability	Interpretability	Computational Cost
Random Forest	High	Moderate	Medium
SVM	Medium	Low	Medium
K-Means	High	Medium	Low
DBSCAN	Medium	Low	Low

Source: Authors, (2026).

The values recorded reflected an inverse relationship between computational cost and interpretability. Random Forest demonstrated greater stability in heterogeneous contexts, while K-Means achieved remarkable efficiency in thematic clustering tasks. This evidence supported the adoption of a mixed approach that combined supervised and unsupervised models to obtain a comprehensive view of the Ecuadorian labor market.

IV.2.2 Findings by Technique Applied

The individual results by technique showed different levels of performance depending on the type of analytical task. As summarized in Table 3, Random Forest outperformed the other techniques in classification metrics, achieving an F1-score of 0.91 and an AUC of 0.93, which demonstrated strong predictive performance. K-Means achieved an internal cohesion value of 0.70, confirming its effectiveness in segmenting thematic profiles, while DBSCAN excelled in detecting anomalies and nonlinear patterns. SVM, although less robust in the presence of noise, proved useful in smaller and highly homogeneous datasets.

Table 3: Comparison between findings by technique.

Technique	Scalability	Computational Cost
Random Forest	F1-score greater than 0.91; AUC > 0.93; strong performance in binary and multi-class classification	Higher computational cost in large datasets; slower than SVM
SVM	Fast training in datasets with fewer than 10,000 records; competitive accuracy	Sensitive to data dispersion and semantic noise
K-Means	Identification of thematic clusters (technological, managerial, operational, and hybrid); silhouette coefficient > 0.70	Required prior dimensional reduction and manual adjustment of the number of clusters
DBSCAN	Detection of nonlinear clusters and outliers	High sensitivity to density parameters and lower stability in scattered data

Source: Authors, (2026).

IV.2.3 Validation of the Job Competency Analysis Model

The validation of the classification models reaffirmed the superiority of Random Forest, which achieved the highest values in precision and sensitivity. Table 4 presented the performance metrics obtained from the processed dataset.

Table 4: Results of the models applied to the identification of competencies.

Model	Accuracy	Precision	Sensitivity (Recall)	F1-score
Random Forest	0.9667	0.9730	0.9600	0.9664
SVM	0.8383	0.8503	0.8383	0.8419

Source: Authors, (2026).

The results indicated that Random Forest achieved an F1-score of 96.64%, significantly outperforming SVM, which reached an F1-score of 84.19%. These metrics confirmed the suitability of Random Forest for identifying complex patterns in heterogeneous labor data, demonstrating superior generalization capability and lower error rates.VLikewise, the unsupervised models K-Means and DBSCAN reinforced the exploratory analysis by identifying four main thematic clusters—technological, managerial, operational, and hybrid—which were validated using a silhouette coefficient of 0.7482. This value demonstrated the internal consistency of the resulting cluster structure.

IV.2.4 Competency Segmentation and Cluster Analysis

Cluster analysis revealed a coherent structure in the organization of professional competencies. The resulting categories combined technical skills—such as programming, data analysis, and visualization—with soft skills including communication, leadership, and customer service. The identified clusters showed the coexistence of technical and soft competencies, indicating the presence of hybrid professional profiles within the analyzed labor market. This pattern was observed consistently across the clustered data, reflecting structural changes in the composition of demanded skills within digitally mediated employment contexts. The segmentation was represented using the elbow method and the silhouette coefficient, as illustrated in Figure 1. The results indicated that the inflection point of the elbow method stabilized at $k = 4$, which coincided with the optimal silhouette coefficient value (≈ 0.7482). This convergence validated the thematic segmentation obtained and confirmed the internal consistency of the clustering model.

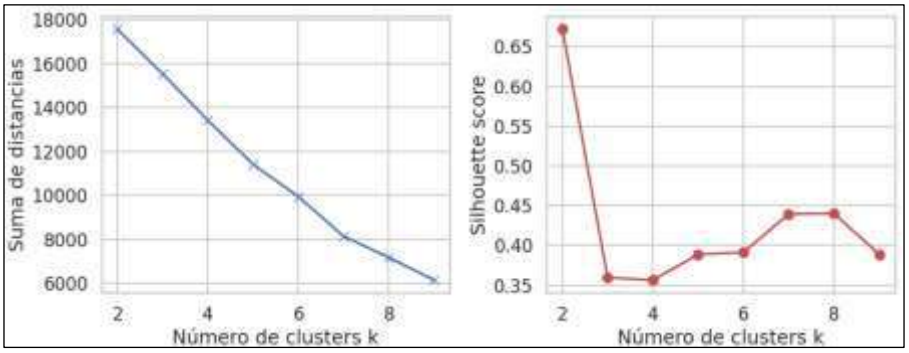


Figure 1: Determination of the optimal number of clusters using the elbow method and silhouette coefficient.
Source: Authors, (2026).

IV.2.5 Classification of Technical and Soft Skills

Semantic classification enabled the clear differentiation between the technical and soft skills most in demand in the Ecuadorian labor market. Skills in Python and R programming, database management, Power BI, and accounting management were identified as highly sought after, along with soft skills such as leadership, teamwork, customer service, and effective communication. The prominence of both technical and soft skills reflected a broader shift observed in digitally mediated labor markets. Digital platforms increasingly prioritized hybrid profiles capable of combining technological proficiency with communication, adaptability, and autonomous problem-solving. This trend was linked to the growing need for workers in platform-based environments to navigate algorithmic interfaces, self-manage tasks, and interact with clients or teams asynchronously.

Therefore, the demand for hybrid competencies was not only an occupational preference but also a structural characteristic of the digital economy, where the boundaries between technical expertise and interpersonal capabilities became progressively blurred. Figure 2 shows the predominance of cross-cutting and operational skills, followed by management and technology skills. These results confirmed the importance of strengthening training programs that integrate both technical and soft components, particularly in sectors undergoing rapid digital transformation, such as technology, commerce, and education.

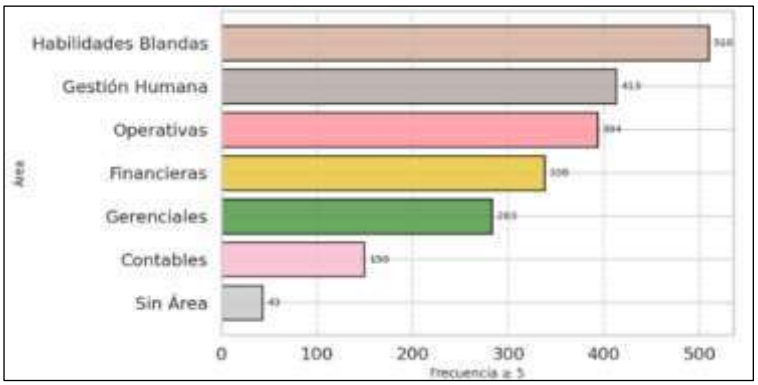


Figure 2: Thematic distribution of standardized competencies extracted from job offers.
Source: Authors, (2026).

IV.2.6 Relative Frequency of Competencies and Territorial Concentration

Frequency analyses generated a word cloud (see Figure 3) that highlighted the most recurrent competencies extracted from job offers. Among the most frequent skills were effective leadership, effective communication, logistics management, financial control, and data analysis, indicating a high demand for multidisciplinary professional profiles.



Figure 3: Word cloud generated from competencies extracted from job offers.
Source: Authors, (2026).

In parallel, the analysis identified a strong geographical concentration of formal employment, as illustrated in Figure 4. Quito (2,328 job offers) and Guayaquil (1,617 job offers) accounted for more than 70% of the total vacancies, followed by Manta (226) and Cuenca (204). The results revealed a pattern of urban concentration accompanied by emerging participation from intermediate cities.

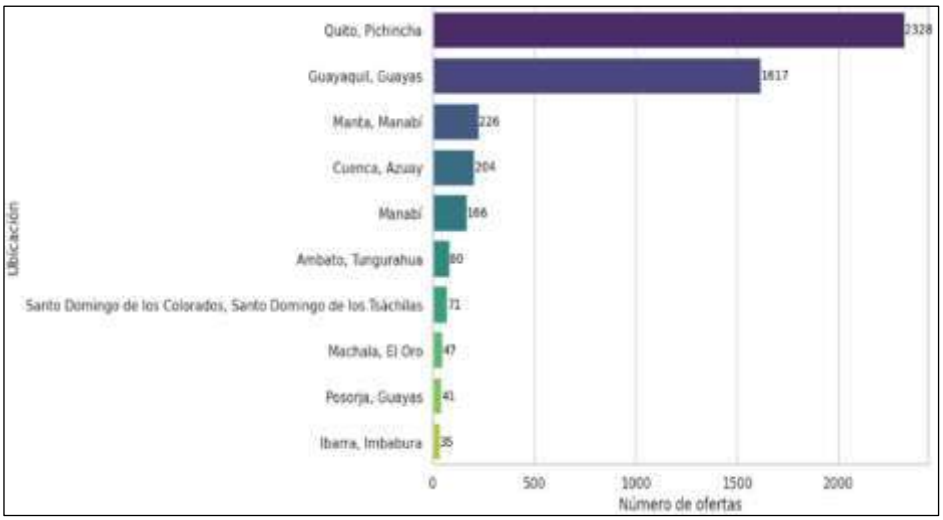


Figure 4: Geographical distribution of job offers in Ecuador.
Source: Authors, (2026).

These geographic disparities reflected structural differences in the spatial distribution of formal employment opportunities. The concentration of vacancies in Quito and Guayaquil was consistent with urban clustering patterns commonly observed in digitally mediated labor markets, where access to connectivity and technological infrastructure was higher than in intermediate or peripheral regions. At the same time, the presence of vacancies in intermediate cities indicated a gradual spatial expansion of formal digital employment.

IV.2.7 Business Concentration in the Labor Market

Figure 5 presented the ten companies with the highest volume of job offers between 2020 and 2024. A high presence of vacancies labeled as Confidential was observed, indicating the widespread use of anonymity practices in digital recruitment processes.

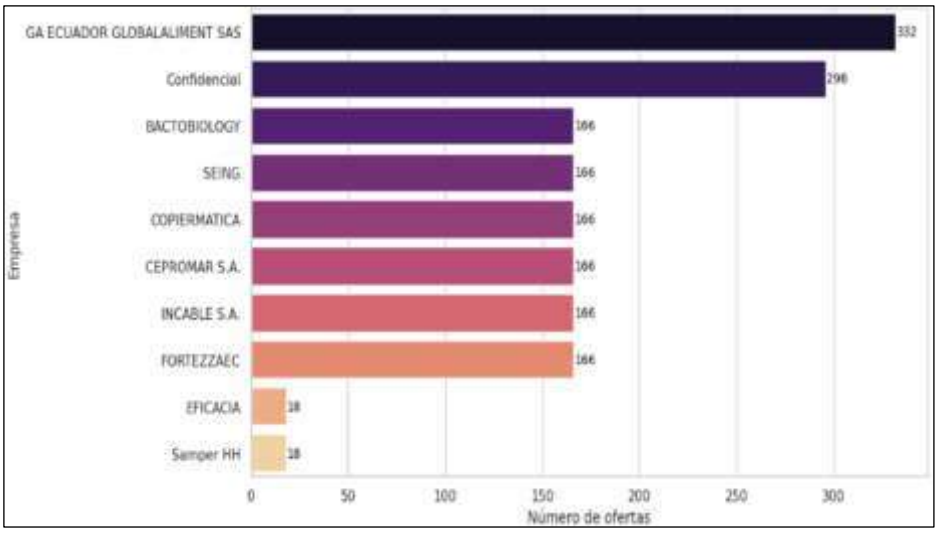


Figure 5: Main companies offering jobs in the Ecuadorian labor market (2020–2024).
Source: Authors, (2026).

The results indicated a high concentration of job offers, in which a limited group of employers generated a substantial proportion of total vacancies. The data showed that large corporations dominated hiring activity during the analyzed period, while medium-sized companies also appeared among the main job offerers, particularly in the logistics and service sectors.

IV.2.8 Model Validation and Comparative Performance

Figures 6 and 7 presented the visual results of the statistical validation performed using confusion matrices and Receiver Operating Characteristic (ROC) curves, respectively. In both analyses, Random Forest exhibited the highest performance, achieving an area under the curve (AUC) value of 0.99, which surpassed the decision tree (AUC = 0.93) and logistic regression (AUC = 0.96).

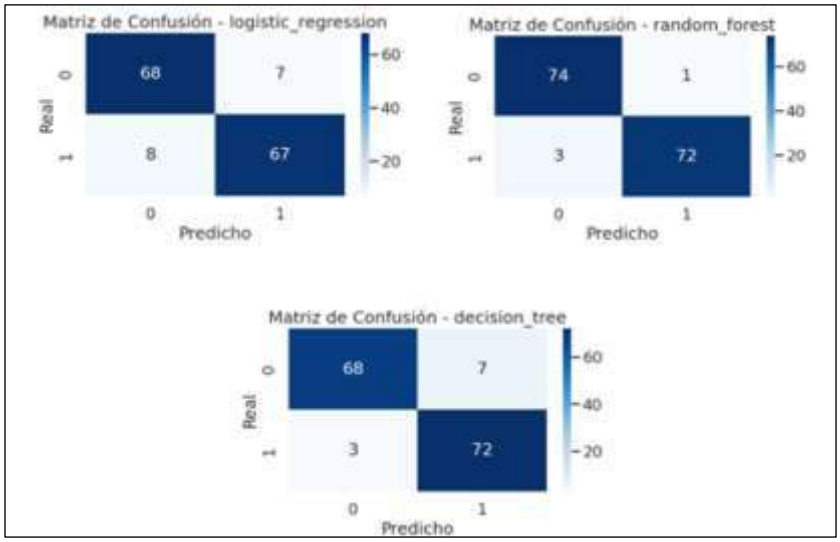


Figure 6: Comparison of confusion matrices between supervised models.
Source: Authors, (2026).

The values shown in Figure 6 indicated that Random Forest achieved 146 correct predictions out of 150 cases, outperforming the decision tree, which recorded 140 correct predictions, and logistic regression, which reached 135 correct predictions in terms of accuracy.

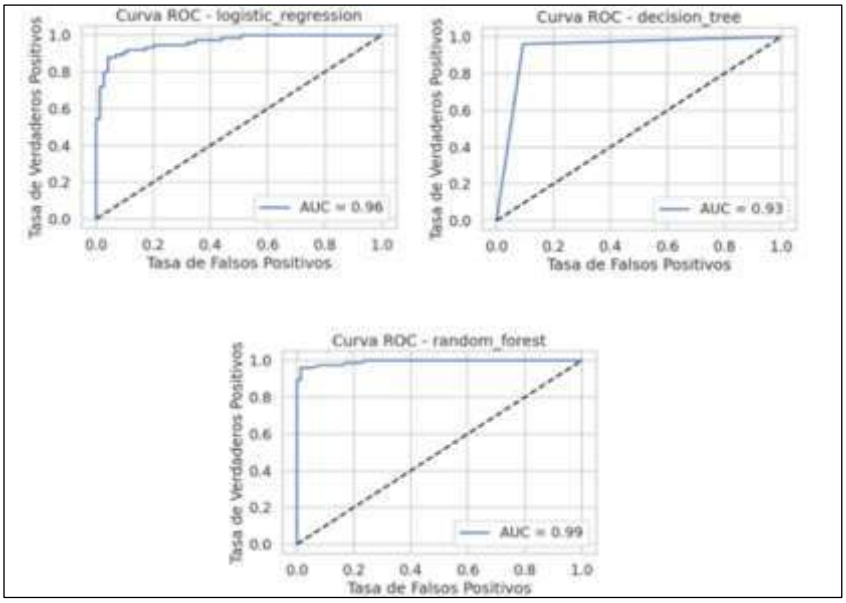


Figure 7: Performance comparison using ROC curves: Logistic Regression, Decision Tree, and Random Forest.
Source: Authors, (2026).

The ROC curves presented in Figure 7 showed consistent performance of the Random Forest model across different classification thresholds, including scenarios characterized by higher semantic variability. These results supported its superior discriminative capacity compared with the other supervised models evaluated.

IV.2.9 Summary of the Comparative Analysis

The results obtained led to the identification of three main findings derived from the comparative analysis. First, the integration of natural language processing with Random Forest provided a favorable balance between accuracy, scalability, and adaptability among the evaluated techniques. Second, the analysis revealed a clear trend toward technical hyper-specialization and the emergence of hybrid professional profiles, accompanied by a strong concentration of job offers in major urban centers. Third, the comparative performance results indicated that hybrid analytical approaches combining semantic analysis with supervised learning achieved higher effectiveness for monitoring professional skill dynamics within the analyzed labor market.

IV.3 DESIGN OF A MODEL FOR DETECTING PROFESSIONAL COMPETENCIES

The analytical model enabled the identification of professional skills in high demand within the Ecuadorian labor market, as well as the hierarchical structure of competencies observed between 2020 and 2024. The results indicated that the integration of structured sources from the National Institute of Statistics and Census (INEC) with unstructured data from digital job portals enhanced the model's predictive performance, achieving accuracy levels above 96%.

IV.3.1 Interactive Visualization and Distribution of Job Offer

The dashboard developed for result visualization integrated geographic, sectoral, and thematic analyses. Figure 8 presented the main interface of the system, which displayed overall dataset metrics, the number of job offers by data source, and dynamic visualizations associated with skills and regions.



Figure 8: Main interface of the dashboard developed in Streamlit.
Source: Authors, (2026).

Figure 8 showed that most job offers originated from private employment portals, with Multitrabajos and BuscoJobs standing out as the sources with the highest volume of vacancies. The indicators revealed a predominance of job offers associated with the commercial, administrative, and technological sectors, together with a clear concentration in the provinces of Pichincha and Guayas.

IV.3.2 Identification of Predominant Skills

The model classified and standardized the competencies identified in thousands of job descriptions, revealing a marked trend toward hybrid skills. Figure 9 presented the main competencies extracted through semantic processing, grouped according to frequency and relevance. The results indicated that the most frequent skills included leadership, customer service, effective communication, logistics management, data analysis, and Python programming, reflecting a convergence between technical and soft competencies within the analyzed dataset.

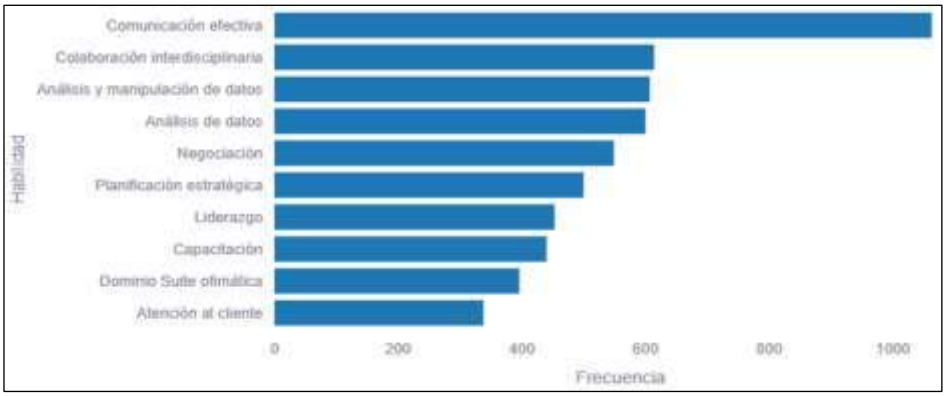


Figure 9: Main skills extracted through semantic processing in Streamlit.
Source: Authors, (2026).

The results showed that job offers across the analyzed sectors frequently required profiles combining technological expertise with management and coordination-related skills. This pattern was observed consistently across all sectors included in the analysis.

IV.3.3 Hierarchical Structure of Competencies

The analysis of semantic hierarchies allowed the representation of the organization of competencies according to levels of generality and frequency. Figure 10 presented the sunburst chart summarizing the relationships between main competencies and sub-competencies derived from the processed textual corpus. The results revealed three dominant hierarchical levels: (i) technological competencies, including data analysis, database management, programming, and Power BI; (ii) management and leadership competencies; and (iii) operational and service-related competencies.

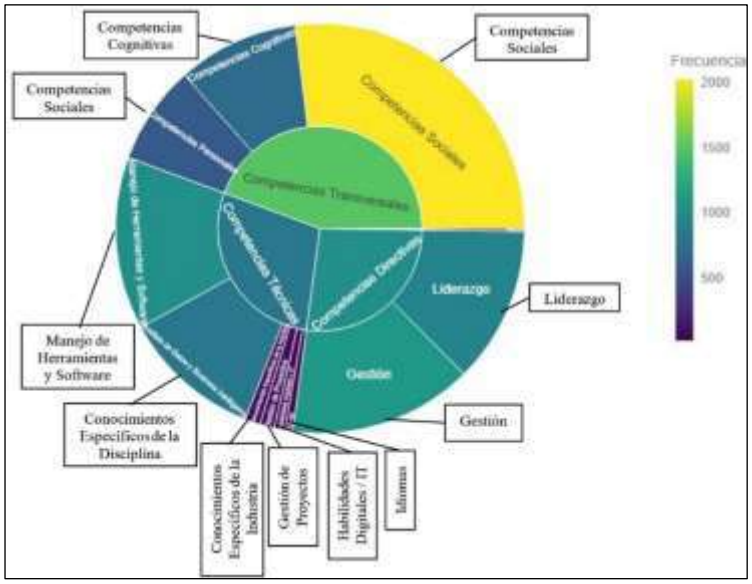


Figure 10: Hierarchy of job skills identified using a sunburst chart.
Source: Authors, (2026).

The technology and commercial sectors exhibited the highest diversity of sub-skills, whereas administrative areas showed lower thematic dispersion. The hierarchical visualization highlighted the interconnection between technical and soft competencies observed in the analyzed data, indicating a structural organization of skills across sectors.

IV.3.4 Model Performance and Consistency of Results

The model's performance showed high accuracy in classifying competencies extracted from unstructured job-related texts. The statistical results indicated an accuracy of 96.67%, a precision of 97.30%, and an F1-score of 96.64%, which validated the robustness of the model under conditions of semantic variability. The consistency observed between the visual outputs and the quantitative performance metrics demonstrated that the model generalized labor demand patterns across different regions of the country.

In addition, the dashboard visualizations enabled the identification of associations between skill types and geographic location. In Quito and Guayaquil, higher demand for technological and administrative profiles was observed, whereas in Manta and Cuenca, operational skills related to logistics and customer service predominated. These spatial patterns were consistently reflected across the visual and quantitative results generated by the system.

IV.3.5 Final Model Architecture

Figure 11 summarized the architecture of the final analytical model, which combined supervised and unsupervised learning techniques within a sequential analysis workflow. The architecture operated through an extraction, transformation, and loading (ETL) process, followed by semantic modeling, statistical validation, and dynamic visualization stages.

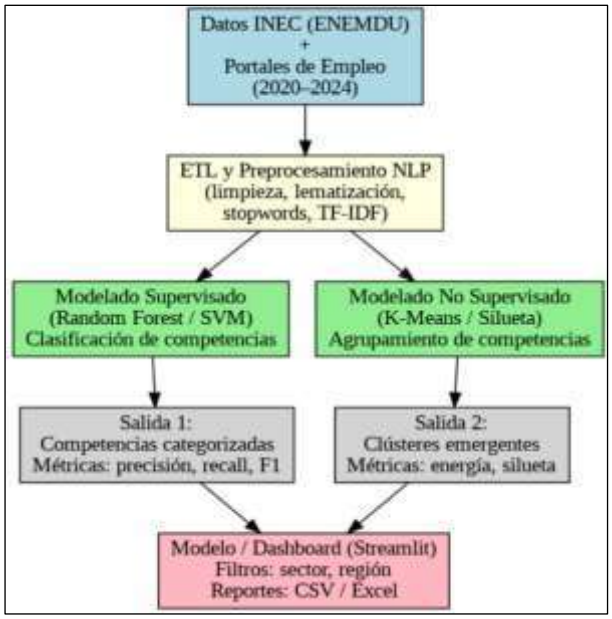


Figure 11: Architecture of the job competency analysis model.
Source: Authors, (2026).

The modular structure allowed algorithms to be updated and new data sources to be incorporated without reconfiguring the overall system architecture, ensuring scalable, transparent, and replicable operation. The resulting configuration supported continuous monitoring of professional competencies based on the analyzed datasets and facilitated the systematic analysis of labor demand patterns within the Ecuadorian context.

IV.4 DISCUSSION

The results of this study demonstrate that applying data science techniques to the analysis of the Ecuadorian labor market is an effective approach for detecting patterns of professional demand and anticipating emerging skills trends. Beyond the technical accuracy achieved—exceeding 96%—the findings show that computational models offer a meaningful way to understand how labor behaviors are increasingly shaped by digital platforms and socio-technical ecosystems. Integrating structured and unstructured sources thus provides not only richer predictive capacity but also insights into how digital employment infrastructures structure access to work opportunities. The hierarchical analysis of competencies revealed a clear trend toward skill hybridization, where technical competencies related to programming, data analytics, and digital management are complemented by soft skills such as leadership, effective communication, and problem solving.

This phenomenon, observed globally in digitally mediated labor markets, reflects the transition toward interdisciplinary profiles capable of operating in environments characterized by automation, virtual collaboration, and accelerated technological cycles. These findings are consistent with research in Latin America [26], [27], which documents a sustained rise in demand for professionals who combine advanced digital abilities with cross-cutting competencies needed for platform-based and remote work environments. Likewise, the results coincide with findings by [28], who emphasize that integrating supervised and unsupervised learning enables deeper and more consistent identification of semantic patterns than traditional statistical methods. This study expands that perspective by incorporating large-scale web scraping and text-mining techniques, enabling the analysis of both formal and informal segments of the labor market.

Such a hybrid methodological approach contributes to addressing a documented gap in the regional literature, revealing how platform-mediated employment and uneven digital adoption reshape skill requirements in heterogeneous work contexts [29]. This broader analytical scope also exposes structural inequalities in skill demand and training gaps, which become more visible when analyzing labor markets through digital traces. The empirical evidence obtained has relevant theoretical and practical implications. Academically, it offers a replicable methodological framework for studies on employability and digital skills, reinforcing the potential of data science for building evidence-based labor observatories. In practical terms, the model provides useful inputs for public policy, curriculum design, and business strategies in human-talent management. When combined with an interactive dashboard, the system enables continuous monitoring of the evolution of labor demand by sector, region, and specialization level.

This capability is crucial for digital-transformation agendas and can support governments in designing e-employment policies, anticipating future skill shortages, and reducing territorial gaps in digital labor inclusion. Nevertheless, some limitations must be acknowledged. Reliance on digital sources may underrepresent sectors characterized by low digitization or high informality, reflecting the broader issue of digital inequality in emerging economies. Moreover, variability in the language of job advertisements can introduce semantic noise, despite the model's strong generalization metrics. Future research could expand temporal coverage, incorporate deep learning architectures to enhance contextual understanding of job descriptions, and integrate socioeconomic, gender, and territorial variables to provide a more comprehensive view of inequality in digital labor markets. Such improvements would strengthen both the predictive capacity and social relevance of the system, enabling its evolution into a full digital labor observatory capable of capturing transformations in work, skills, and employment in digitally evolving societies.

V. CONCLUSIONS

This study demonstrated that the application of data science techniques constitutes an effective approach for analyzing and understanding the Ecuadorian labor market in a context increasingly mediated by digital technologies. By integrating structured sources, such as INEC records, with unstructured data from digital job platforms, the proposed model accurately identified professional competencies, revealed training gaps across productive sectors, and captured patterns of hybrid skill demand combining technical and soft skills. The high performance achieved—exceeding 96% accuracy—confirmed the value of combining supervised and unsupervised learning within a modular and scalable analytical framework supported by natural language processing and interactive visualizations. Although reliance on digital job sources may underrepresent informal or low-digitization sectors and introduce semantic variability, the model showed consistent and robust performance. Overall, this research contributes a replicable methodological approach to digital labor analytics and provides a foundation for evidence-based decision making, workforce planning, and future extensions aimed at monitoring emerging skills and addressing territorial and socio-digital inequalities in evolving labor ecosystems.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres
Methodology: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres
Investigation: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres
Discussion of results: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres
Writing – Original Draft: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres
Writing – Review and Editing: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres
Resources: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres
Supervision: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres
Approval of the final text: Llerena Lucrecia, Rodríguez Nancy, Bustamante Jefferson and Angel Torres

VII. REFERENCES

[1] L. Bareño Pinzón, P. Orjuela Garzón, and F. Pinzón Galeano, "Análisis de los beneficios de las nuevas tecnologías implementadas al proceso de reclutamiento de personal a partir del Covid19 en grandes empresas colombianas en el 2023," Universitaria Agustiniana, Bogotá D. C., 2023. Accessed: Aug. 05, 2025. [Online]. Available: http://repositorio.uniagustiniana.edu.co/handle/123456789/2496

[2] A. Taboada Villamarín, "Big data en ciencias sociales. Una introducción a la automatización de análisis de datos de texto mediante procesamiento de lenguaje natural y aprendizaje automático," *Revista CENTRA de Ciencias Sociales*, vol. 3, no. 1, Jan. 2024, doi: 10.54790/rccs.51.

[3] F. Da Silva and N. Georgina, "La era de las plataformas digitales y el desarrollo de los mercados de datos en un contexto de libre competencia," Documentos de Proyectos (LC/TS.2021/173) Dec 2021, Santiago, Comisión Económica para América Latina y el Caribe (CEPAL). Accessed: Sep. 08, 2024. [Online]. Available: https://www.cepal.org/es/publicaciones/47540-la-era-plataformas-digitales-desarrollo-mercados-datos-un-contexto-libre

[4] M. Jiang and P. Jia, "Does the level of digitalized service drive the global export of digital service trade? Evidence from global perspective," *Telematics and Informatics*, vol. 72, p. 101853, Aug. 2022, doi: 10.1016/j.tele.2022.101853.

[5] A. Alibasic, H. Upadhyay, M. C. E. Simsekler, T. Kurfess, W. L. Woon, and M. A. Omar, "Evaluation of the trends in jobs and skill-sets using data analytics: a case study," *J Big Data*, vol. 9, no. 1, p. 32, Dec. 2022, doi: 10.1186/s40537-022-00576-5.

[6] R. Rodríguez, "Brecha digital y transformación social: el impacto de las nuevas tecnologías en América Latina y el Caribe," *Revistas UPR*, vol. 5, 2024.

[7] ILO, "Progress in the labour market in Latin America and the Caribbean remains insufficient," International Labour Organization. Accessed: Jan. 05, 2025. [Online]. Available: https://www.ilo.org/resource/news/insufficient-labour-market-progress-latin-america-caribbean-2024-overview-report?utm_source=chatgpt.com

[8] IDB, "Using Big Data and its Analytical Techniques for Public Policy Design and Implementation in Latin America and the Caribbean," Inter-American Development Bank. Accessed: Jan. 09, 2025. [Online]. Available: https://publications.iadb.org/publications/english/document/Using-Big-Data-and-its-Analytical-Techniques-for-Public-Policy-Design-and-Implementation-in-Latin-America-and-the-Caribbean.pdf

[9] M. Villamar Choez, G. Mite Vivero, M. Vera Franco, and G. Figueroa Morán, "Impacto de TIC en el mercado laboral aplicado a egresados, carrera tecnologías de la información," *Journal TechInnovation*, vol. 3, no. 0, pp. 40–46, Dec. 2024.

[10] INEC, "Metodología del Registro Estadístico de Empleo en la Seguridad Social (REESS)," Instituto Nacional de Estadística y Censos.

[11] A. Kouskoura, E. Kalliontzi, D. Skalkos, and I. Bakouros, "Evaluating Experts' Perceptions on Regional Competitiveness Based on the Ten Key Factors of Assessment," *Sustainability*, vol. 16, no. 14, p. 5944, Jul. 2024, doi: 10.3390/su16145944.

[12] L. Echeverría and G. Rucci, "¿Qué suma la ciencia de datos a la identificación y anticipación de la demanda de habilidades?," Washington, D. C., Nov. 2022. doi: 10.18235/0004554.

[13] O. Grljević and T. Kecojević, "An NLP approach to skill analysis in ICT job advertisements from a gender perspective," *Strategic Management*, no. 00, pp. 94–94, 2025, doi: 10.5937/StraMan2500008G.

[14] V. Escudero, H. Liepman, and A. Podjanin, *Big Data Applications in Labor Economics, Part B*. Emerald Publishing Limited, 2024. doi: 10.1108/S0147-9121202552B.

[15] O. Kässi and V. Lehdonvirta, "Online labour index: Measuring the online gig economy for policy and research," *Technol Forecast Soc Change*, vol. 137, pp. 241–248, Dec. 2018, doi: 10.1016/j.techfore.2018.07.056.

[16] C. Muntaner, "Digital Platforms, Gig Economy, Precarious Employment, and the Invisible Hand of Social Class," *International Journal of Health Services*, vol. 48, no. 4, pp. 597–600, Oct. 2018, doi: 10.1177/0020731418801413.

[17] K. M. Kuhn, J. Meijerink, and A. Keegan, "Human Resource Management and the Gig Economy: Challenges and Opportunities at the Intersection between Organizational HR Decision-Makers and Digital Labor Platforms," 2021, pp. 1–46. doi: 10.1108/S0742-730120210000039001.

[18] D. Autor and A. Salomons, "Is automation labor displacing? Productivity growth and the labor share," *Brookings Pap Econ Act*, vol. 53, no. 2, pp. 1–58, 2018, Accessed: Mar. 07, 2025. [Online]. Available: https://www.brookings.edu/wp-content/uploads/2018/03/1_autorsalomons.pdf

[19] A. Henry and R. Essah, "Artificial Intelligence and Workforce Transformation: An Analysis of Employment Disruption and Adaptation Strategies," *International Journal of Latest Technology in Engineering Management & Applied Science*, vol. 14, no. 8, pp. 493–501, Sep. 2025, doi: 10.51583/IJLTEMAS.2025.1408000059.

[20] H. Kleven, C. Landais, and J. E. Søgaard, "Children and Gender Inequality: Evidence from Denmark," *Am Econ J Appl Econ*, vol. 11, no. 4, pp. 181–209, Oct. 2019, doi: 10.1257/app.20180010.

[21] E. Struffolino and M. Raitano, "Early-Career Complexity Before and After Labour-Market Deregulation in Italy: Heterogeneity by Gender and Socio-economic Status Across Cohorts," *Soc Indic Res*, vol. 151, no. 1, pp. 231–257, Aug. 2020, doi: 10.1007/s11205-020-02373-0.

[22] J. A. Toledo Íñiguez and D. A. Angulo Vélez, "Automatización de procesos administrativos en instituciones públicas en Ecuador," *Revista Científica Multidisciplinar G-nerando*, vol. 6, no. 1, Mar. 2025, doi: 10.60100/rcmg.v6i1.547.

[23] M. G. Mayor Ravines, J. G. Mayor Gamero, and J. V. Mayuri Barrón, "Transformación digital en la administración pública: una revisión sistemática de marcos, desafíos y resultados (2013–2023)," *Nexus Research Journal*, vol. 4, no. 2, pp. 35–63, Aug. 2025, doi: 10.62943/nrj.v4n2.2025.325.

[24] A. Aristovnik, D. Ravšelj, and E. Murko, "Decoding the Digital Landscape: An Empirically Validated Model for Assessing Digitalisation across Public Administration Levels," *Adm Sci*, vol. 14, no. 3, p. 41, Feb. 2024, doi: 10.3390/admsci14030041.

[25] T. Bielialov, I. Kalina, V. Goi, O. Kravchenko, N. Shyshpanova, and A. Negoda, "Global Experience of Digitalization of Economic Processes in the Context of Transformation," *Journal of Law and Sustainable Development*, vol. 11, no. 3, p. e814, Aug. 2023, doi: 10.55908/sdgs.v11i3.814.

[26] A. Zuñiga-Collazos, J. M. Velásquez Orozco, and A. Rojas-Ospina, "Digital Competences and Their Impact on Employability in the Tourism Sector—An Applied Study," *Sustainability*, vol. 17, no. 13, p. 6133, Jul. 2025, doi: 10.3390/su17136133.

[27] C. Buenadicha Sánchez *et al.*, "21st Century Skills: Transversal Skills Development in Latin America and the Caribbean," Washington, D. C., Mar. 2019. doi: 10.18235/0001574.

[28] C. Quan, M. Wang, and F. Ren, "An Unsupervised Text Mining Method for Relation Extraction from Biomedical Literature," *PLoS One*, vol. 9, no. 7, p. e102039, Jul. 2014, doi: 10.1371/journal.pone.0102039.

[29] J. Schedlbauer, G. Raptis, and B. Ludwig, "Medical informatics labor market analysis using web crawling, web scraping, and text mining," *Int J Med Inform*, vol. 150, p. 104453, Jun. 2021, doi: 10.1016/j.ijmedinf.2021.104453.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 964-975. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3141

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

PERFORMANCE AND EMISSION OPTIMIZATION OF A DIESEL ENGINE USING KARANJA BIODIESEL–HYDROGEN DUAL-FUEL WITH AZADIRACHTA INDICA ANTIOXIDANT

Kumaran P*[1], Natarajan S[2] , Sathiyaraj S[3], Shanmugasundaram K[4] and Subramanian T[5]

[1, 3,4,5]Department Mechanical Engineering, Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research Foundation(deemed to be university), Salem, Tamilnadu, India.

[2]Department Mechanical Engineering, Vinayaka Mission's Kirupananda Variyar Engineering College, Vinayaka Mission's Research Foundation(deemed to be university), Salem, Tamilnadu, India.

[1]http://orcid.org/ 0000-0003-2781-8999, [2]http://orcid.org/0000-0003-4601-2800, [3]http://orcid.org/0000-0001-8504-0412
[4]http://orcid.org/0009-0005-8298-5468, [5]http://orcid.org/0009-0000-9836-1944

Email: *kumaranp@avit.ac.in, natarajanshree@gmail.com, sathiyaraj@avit.ac.in, kssanjay.8@gmail.com and mani.thiyagu2308@gmail.com

ARTICLE INFO

Article History
Received: December 15, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Renewable fuels,
Internal combustion engines,
Environmental degradation,
Fossil fuel depletion,
Hydrogen fuel,
Dual-fuel mode

ABSTRACT

Renewable fuels are critically required for internal combustion engines because of environmental degradation and the exhaustion of fossil fuels. Hydrogen, possessing a higher auto-ignition temperature and zero emissions has recently surfaced as a feasible alternative to biodiesel and other renewable fuels. Integrating hydrogen with an alternative liquid fuel and operating the engine in dual-fuel mode reduces the self-ignition temperature of the compression ignition engine. This study aims to investigate the impact of including a natural antioxidant on the performance of a diesel engine operating on biodiesel with an elevated hydrogen concentration. During the testing method, hydrogen is injected into the system by the intake manifold, while a B15 mixture of Karanja biodiesel is delivered into the combustion chamber. The experimental fuel samples comprised Diesel, B15, and B15 augmented with hydrogen at flow rates of 6, 9, 12, 15, and 18 lpm. The combination of B15 with hydrogen results in an enhancement of brake thermal efficiency (BTE). Blends of B15 with 15 liters of gasoline and B15 with 18 liters of gasoline demonstrated superior performance compared to the traditional engine at full load CO, achieving BTE improvements of 28.60% and 27.62%, respectively. HC emissions from B15 at 15 lpm and B15 at 18 lpm are reduced in comparison to other mixtures, however NOx emissions from diesel are elevated by 25 and 27 %, respectively. The combination of B15 + 15 lpm with 1000 ppm of Azadirachta indica (AI), a natural antioxidant, diminished NOx emissions to 1074 ppm, reflecting a reduction of 21.9% relative to B15 + 15 lpm without AI. The outcome was a 2.17% enhancement in BTE relative to B15 + 15 lpm without AI. Consequently, numerous test mixes are inferior than B15 + 15lpm with additive.

Copyright ©2025 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

1. INTRODUCTION

Growing environmental concerns and the depletion of fossil fuels have accelerated the search for cleaner and more sustainable alternatives for compression ignition engines. Dual-fuel operation, combining biodiesel with hydrogen, offers the potential to enhance combustion efficiency while reducing harmful emissions. The addition of natural antioxidant additives further improves fuel stability and performance characteristics. Optimization of such blends enables a balanced approach to achieving higher efficiency with lower environmental impact. L-ascorbic acid (LA) antioxidant effects were explored in a lanthanum oxide-coated low heat rejection engine using AME, WCOME, and JME biodiesel blends.

AME20 recorded the highest BTE (30.3%) with lower HC, CO, and smoke emissions than diesel, though NOx slightly increased (2.1%). Adding LA200 mg further reduced HC by 15.8%, CO by 10.4%, and smoke by 9.3%, indicating antioxidant-assisted blends can reduce oxidation-induced NOx without major efficiency losses [1]. Natural antioxidants (NA) explored from Albizia lebbeck leaves in an LHR engine (zirconia-coated piston and head) using SBME20 (80% diesel + 20% sesame biodiesel). NA addition reduced BSFC to 0.38 kg/kW-hr (from 0.42 in CE), CO to 0.2 g/kWh, and NOx to 0.5 g/kWh at higher loads. Smoke opacity fell by 10%, showing NA's role in stabilizing biodiesel and improving combustion in high-temperature LHR conditions [2]. This study examined the addition of diphenylamine (DPA) and ceria nanoparticles (CeO_2) to a B30 Jatropha biodiesel blend. The optimal B30+DPA50+$CeO_2$50 blend achieved a 6.35% reduction in brake-specific fuel consumption (BSFC) and an 8.68% decrease in NOx emissions, with a 5.74% increase in brake thermal efficiency (BTE). Machine learning modelling showed that the K-nearest neighbors algorithm provided the most accurate predictions compared to multilayer perceptron and random forest regression [3]. ANN and Taguchi-based optimization confirmed improved stress transfer and eco-friendly applicability for packaging and building materials [4]. This study evaluated N-isopropyl-N′-phenyl-1,4-phenylenediamine (IPPD) as an antioxidant in Jatropha methyl ester (JME) for thermal barrier coated (TBC) diesel engines. The piston crown was coated with YSZ–CeO_2 (0.3 mm) to improve thermal efficiency. The optimal 1000 ppm IPPD-doped JME reduced NO emissions by 8.9% while increasing BTE by 18.8% and reducing BSFC by 11.7%. HC, CO, and smoke emissions were also reduced by over 22% at maximum load compared to the JME-fueled conventional engine [5]. Palm biodiesel blends with 10% dimethyl carbonate (DMC) were prepared and tested in a CI engine across speeds from 1100 to 2100 rpm.

DMC addition improved brake power (up to +1.70%) and BTE (up to +4.90%), while reducing BSFC (up to −2.93%) and EGT (up to −4.54%). Emissions decreased significantly, with CO reduced by up to 26.47% and HC by up to 33.33% compared to biodiesel blends without antioxidant [6]. Lemongrass oil (5–20 mL/L) was used as an antioxidant in B20 Calophyllum inophyllum methyl ester (CIME) blends to improve oxidation and thermal stability. The best blend (B20LG20) extended the induction period to 9.6 h, reduced peroxide value by 51%, and limited acid number growth by 42%. Engine tests showed BTE of 31.2%, with CO and smoke emissions reduced by 15% and 25%, and NO reduced by 6% compared to B20 [7].Hybrid composites of Holoptelea integrifolia bark fiber and Ziziphus jujuba seed particles were developed using a tannic acid-based bio-epoxy resin (BBER). Mechanical testing data were modeled using fuzzy logic, achieving an 87% prediction accuracy for tensile, flexural, and impact strengths. The composites demonstrated strong mechanical performance potential with renewable resin systems for sustainable structural applications [8]. Four antioxidants—BHT, BHA (phenols), and DPPD, PPD (aromatic amines)—were tested in neat sunflower biodiesel and B20 blends. PPD showed the greatest NOx reduction (52%) compared to other antioxidants. Adding 1000 ppm antioxidants to B20 reduced NOx more effectively than in neat biodiesel. CO and HC emissions dropped slightly, while smoke number increased with antioxidant concentration and load [9]. Biodiesel was extracted from Scenedesmus obliquus (S1) and Scenedesmus dimorphus (S2) using Ni/H_2 transesterification with a Ni(II)–Schiff base promoter. Blends (S1B10–S1B50, S2B10–S2B50) showed slightly higher SFC and lower brake power but reduced greenhouse gas emissions, except for NOx. The fuels also exhibited antimicrobial, antioxidant activity, and improved thermal and corrosion resistance [10]. Waste cooking oil biodiesel (B20) was enriched with hydrogen (10 L/min) and synthetic antioxidant TBHQ (1000–2500 ppm).

The B20 + 2000 ppm + H_2 blend achieved the highest oxidation stability, 7.39% higher BTE, and 11% lower BSFC than diesel, while reducing NOx by 3.63%. Hydrogen-antioxidant synergy improved performance but lowered ICP and HRR [11]. Diphenylamine (DPA, 100 ppm) was added to B30 Jatropha biodiesel, with performance modeled using RSM and ANFIS. B30 + DPA100 reduced NOx by up to 15.53% compared to B30, with minor changes in BTE and BSFC. BMEP increased by 1.3% over B30, indicating feasibility for use without engine modification [12]. AA2024 metal matrix composites reinforced with nano-ZrO_2 were produced by squeeze casting, with process parameters (melt/die temperatures, pressure) optimized for mechanical and wear performance. The study applied TOPSIS-EWM to identify ideal electrical discharge machining (EDM) settings, ensuring minimal nanoparticle damage and consistent dimensional accuracy. Findings support precise process control for high-quality MMC fabrication [13]. Orange peel methyl ester (OPME) biodiesel was doped with butylated hydroxytoluene (250–1000 ppm) and tested in both conventional and YSZ-coated low-heat rejection engines. At 750 ppm BH, BTE increased by 3.04%, BSFC dropped by 15.3%, and CO and HC fell by 36.6% and 39.1%. Antioxidant addition improved efficiency and lowered emissions [14]. Pongamia methyl ester (PME20) biodiesel was combined with natural antioxidant from Mangifera indica leaves (1000 ppm) and hydrogen enrichment (16 L/min). PME20 + 16lpm + NA1000 achieved 29.17% BTE and 13.5% lower NOx than PME20 + 16lpm. Hydrogen improved efficiency but increased NOx, which was mitigated by the antioxidant [15]. This review highlights the role of additives, especially antioxidants in improving biodiesel stability, combustion, and emission characteristics. Antioxidants enhance storage life, reduce corrosion, and maintain engine cleanliness.

Metal-based additives also aid emission control, making biodiesel more viable for CI engine applications [16]. A low heat rejection (LHR) diesel engine with a TBC piston coated in YSZ and CeO_2 (0.3 mm) was fueled with Jatropha Methyl Ester (JME20) doped with 500–2000 ppm N,N′-diphenyl-p-phenylenediamine (DPPD). The coated engine (CE) running on JME20B4 reduced NO by 6.5%, HC by 17.4%, and CO by 34.6% at full load. Smoke opacity also dropped by 15.3%, though peak HRR and cylinder pressure decreased by 5–6% [17]. Clove (CEO), thyme essential oils (TEO), and synthetic BHT (3000 ppm) were added to B20 colza biodiesel–diesel blends. While antioxidants slightly raised SFC (0.9–2.4%) and reduced thermal efficiency (1.1–2.2%), BHT cut NO emissions by 21% but increased CO by 3.68%. CEO improved overall performance and emissions more than TEO, while all antioxidants reduced smoke and CO_2 [18].Karanja oil biodiesel B20 was blended with carbon nanotubes (CNTs) using QPAN80 surfactant (1:4 NP:surfactant ratio). The B20CNT50 blend boosted BTE by 9.8% and reduced BSFC by 18.7%. Emissions reduced significantly—CO by 7.7%, CO_2 by 11.2%, HC by 10.2%, smoke by 5.7%, and NOx by 10.9%—while machine learning models (SVR, RF) accurately predicted performance and emissions [19]. Karanja oil biodiesel was produced using cobalt-doped zinc oxide (CZO) nanoparticles (~75 nm) by a dual-step esterification–transesterification process. At optimal conditions (1:25 oil:methanol, 0.3 wt% catalyst, 70°C, 180 min), biodiesel yield reached 98.5%. CZO provided high purity, easy glycerol separation, and reusability, making it a commercially viable catalyst for biodiesel production [20]. A single-zone, zero-dimensional progressive combustion model was developed in Python to predict CI engine performance, combustion, and emissions. Validated against experimental data for diesel, Pongamia biodiesel, and blends, the model accurately estimated in-cylinder pressure and HRR.

Results showed blending up to 30% biodiesel had minimal performance impact, though exhaust oxygen concentration increased with blend ratio [21]. Pongamia biodiesel blends (20BD, 50BD, 100BD) were tested in a CI engine with and without thermal barrier coatings (TBC) of NiCrAl (1 mm) and TiO_2 (4 mm). With 20BD in the coated engine, fuel consumption dropped by 10.6% and BTE increased by 4.7% compared to the uncoated engine. Enhanced heat retention improved fuel vaporization, shortened ignition delay, and increased mechanical efficiency [22]. B20 biodiesel blends from Pongamia, Juliflora, and Calophyllum were evaluated in a low heat rejection engine with zirconia-coated components. Adding 2.5% neem leaf extract improved BTE and reduced NOx compared to blends without additives. Biodiesel blends lowered CO, HC, and PM emissions, while increasing cylinder pressure and HRR. Neem extract further minimized ignition delay and enhanced combustion stability [23]. This study presents the findings of a CI engine utilizing Karanja B15 biodiesel-hydrogen in dual fuel mode, enhanced with a natural antioxidant additive, focusing on performance and emission characteristics. The incorporation of the additive into the biodiesel-enhanced hydrogen blend resulted in an elevation of BTE and a reduction in NOx emissions. In addition to diesel engines, this study evaluates the performance and emission parameters of CI engines that use other fuels. This study does not require engine modifications; instead, it necessitates a specialized fuel system for hydrogen fuel, enabling the gas to reach the intake manifold in conjunction with air.

II. MATERIALS AND METHODS

II.1 PREPARATION OF PRIMARY FUEL

The present study utilizes a mechanical crusher to extract Karanja oil, which functions as a feedstock. Prior to the addition of the methoxide solution, 500 cc of Karanja oil is heated to 105 °C. The catalyst in a 150 ml methoxide solution comprises 1.5 wt% potassium hydroxide (KOH). The solution mixture is constantly stirred at 700 rpm using a mechanical stirrer, and the reaction temperature is kept at 60 °C. Subsequent to transferring the liquid to the conical beaker, allow it to settle for a duration of 15 to 24 hours. Subsequently, biodiesel will be observed floating atop, while glycerol will settle at the bottom. The resulting solution is rinsed to remove any excess water, alcohol, or catalyst. Table 3 presents the fuel characteristics of the B100 and B15 Karanja biodiesel blends. The B15 mix resembles diesel fuel more closely and has superior performance compared to the B100 blend.

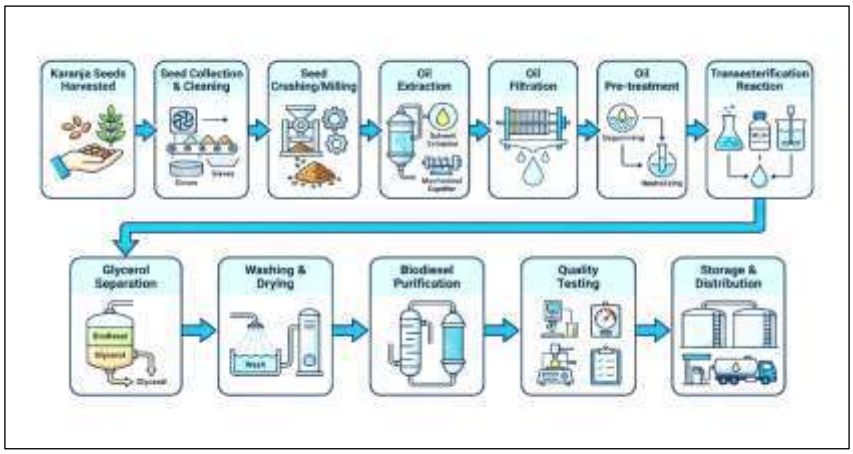


Figure 1: Procedure for making biodiesel.
Source: Authors, (2026).

Table 1: Properties of Test fuel.

Property	Diesel	Karanja Methyl ester	
		B15	B100
Viscosity	2.85	3.27	4.58
Density	832	842	890
Heating Value (kJ/kg)	43,500	42,050	37,050
Cetane index	47	49	53
Flash Point (˚C)	57	63	92
Fire point (˚C)	66	73	100

Source: Authors, (2026).

II.2 PILOT FUEL PREPARATION

Table 2: Properties of hydrogen and diesel.

Parameter	Hydrogen	Diesel
Formula	H_2	$C_n\ h_{1.8*n-c8-c20}$
Heating value (MJ/kg)	141.9	44.8
Self-ignition temperature (K)	773–858	483–573
Viscosity at 300 K (mm^2/s)	95–120	0.4–1.0
Cetane number	-	45–52
Specific Gravity	0.069–0.073	0.82–0.86
Flame Velocity (cm/s)	250–330	25–40

Source: Authors, (2026).

Table 2 enumerates the attributes of diesel and hydrogen. The data in the table indicates that hydrogen possesses a higher self-ignition temperature compared to diesel. This indicates that for optimal combustion, hydrogen fuel ought to be combined with diesel, biodiesel, or ethanol, all of which possess lower ignition temperatures. Using 99.9 % pure hydrogen, experiments are carried out using a modified CI engine that has a flame arrester and other devices. The section on experiments delineates the specific steps.

II.3 ANTIOXIDANT ADDITIVE PREPARATION

The foliage of Azadirachta indica (neem), a tree typically reaching 20–30 m in height (occasionally up to 40 m) was selected for antioxidant extraction owing to its high carotenoid and chlorophyll content, proven radical-scavenging properties, low cost, and wide availability. Neem leaves are pinnate (20–40 cm long) with serrated leaflets, exhibiting lighter green on the lower leaflets and deeper green on the upper ones; the tree produces oval drupes that ripen to a yellowish-green color. Freshly harvested leaves were shade-dried for 10 days to minimize photooxidative degradation, followed by oven-drying at 50–70 °C for 10–15 min to remove residual moisture. To inactivate oxidative enzymes such as polyphenol oxidase, the dried leaves were immersed in water at 50 °C for 10 min. The material was then cut into 1–2 mm fragments to increase surface area and pulverized into a fine powder using a mechanical crusher. For solvent extraction, 100 g of powder was combined with 20 mL of an acetone–ethanol mixture (v/v) and agitated at ambient temperature for 1 h, producing a pale green opalescent suspension. The mixture was allowed to rest for 1 h to facilitate sedimentation, and the supernatant was collected and concentrated under reduced pressure using a rotary evaporator at 40 °C. The concentrate was dried at the same temperature for 72 h to obtain the crude antioxidant extract. This extract was incorporated into a B15 Karanja biodiesel blend at a concentration of 1000 ppm for subsequent oxidative stability and performance evaluations.

II.4 EXPERIMENTATION

This research utilizes a water-cooled, single-cylinder direct injection engine together with all necessary loading and measuring apparatus. Operating conditions can be regulated by sustaining specific ranges for exhaust gas temperature, cylinder pressure, air flow, and fuel flow,. A piezoelectric transducer is utilized to monitor the engine cylinder pressure. The engine possesses a rated power of 3.7 kW, a compression ratio of 16.5:1, and operates at a speed of 1500 rpm. All potential loads, ranging from 0% to 100%, including 25%, 50%, 75%, and full load, are evaluated during the trials. An eddy current dynamometer is employed to modify the load during experiments. A gas analyzer, an AVL smoke meter, and thermocouples are employed to measure emissions including hydrocarbons, smoke, and nitrogen oxides, in addition to the exhaust temperature and fuel flow rate. A flow meter and flame arrester are used to feed more hydrogen into the CI engine's air and hydrogen intake systems. Flame arrestors are a group of devices used in hydrogen systems that are intended to stop explosions. Flame traps have proven effective in lowering backfire and other unfavorable combustion events in a hydrogen engine system.

A non-return valve (NRV) in the fuel line prevents gases from returning to the cylinder. Hydrogen is supplied by a 140-bar high-pressure cylinder with an outlet pressure of 2 bar. Upon traversing the flame arrester and trapper, the hydrogen proceeds to the non-return control valve. The digital flow meter quantifies the flow rate. As additional hydrogen gas reaches the non-return valve, it inhibits reverse flow, directing the gas to the engine. The flame arrestor is then filled with hydrogen. A flame arrestor can prevent an explosion in the fuel encasing system. The fundamental concept of the flame arrestor is that the gas must provide sufficient heat to sustain the flame. The enrichment of hydrogen transpires when air and hydrogen are amalgamated post-collection at the flame trap unit and subsequently discharged into the input manifold. A variety of hydrogen flow rates—8 lpm, 10 lpm, 12 lpm, 15 lpm, and 18 lpm—were supplied into the input manifold. The findings make it possible to compare the performance and emission characteristics of the CSI engine using a hydrogen and B15 combination with and without the antioxidant AI. Figure 2 illustrates the configuration of the CI engine employed in the trials.

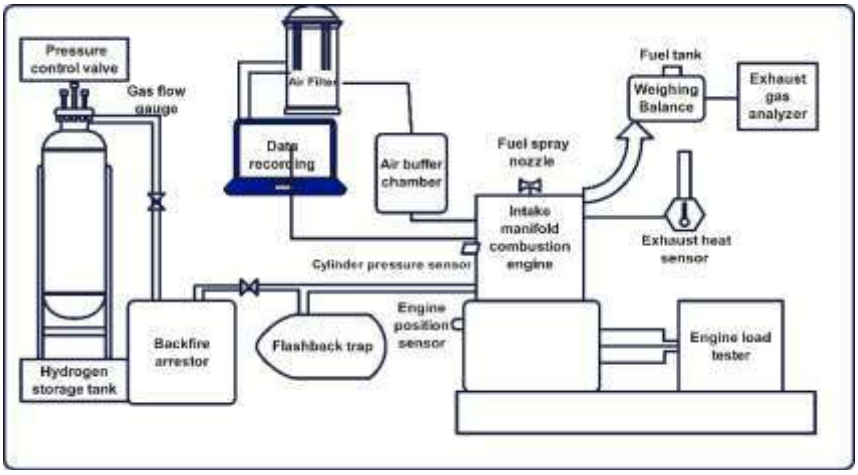


Figure 2: Configuration of CI engine test setup.
Source: Authors, (2026).

The Kirloskar AV-1 is a water-cooled, four-stroke, single-cylinder diesel engine operating continuously at 1500 rpm in the experimental configuration. The engine parameters comprise a 560-cc displacement, 80 mm bore, 110 mm stroke, and a compression ratio of 16.5:1. The injection pressure is maintained at 20 Pa, with the injection duration for both diesel and biodiesel set at 23° CA before top dead center (bTDC). The engine, connected to an eddy current dynamometer for load assessment, can deliver a maximum power output of 3.7 kW.

II.5 UNCERTAINTY ANALYSIS.

The standard deviations and uncertainties of the experimental measurements were subjected to statistical analysis. The fuel volume flow rate, defined as the duration required to consume 10 cc of gasoline, exhibited an absolute error of ±0.1 s and a relative uncertainty of ±0.5%, with a standard deviation of 0.18 s. The overall fuel consumption exhibited a relative uncertainty of ±0.5% and an absolute error of ±0.005 kg/h, with a standard deviation of 0.01 kg/h. The peak pressure exhibited a relative error of ±1%, an absolute uncertainty of ±1 bar, and a standard deviation of 1.58 bar. The exhaust gas temperature fluctuated by 7.97 °C, with absolute uncertainties of ±3.6 °C and relative uncertainties of ±1.1%. The standard deviation for nitrogen oxides (NOx) was 8.72 ppm, accompanied by a relative uncertainty of ±0.5% and an absolute inaccuracy of ±4 ppm. The hydrocarbon emissions exhibited a relative uncertainty of ±1.5%, an absolute error of ±2.5 ppm, and a standard deviation of 5.4 ppm. The carbon monoxide values exhibit a relative inaccuracy of ±5%, an absolute uncertainty of ±0.008%, and a standard deviation of 0.02%.

II.6 ERROR ANALYSIS

It is infeasible to entirely eradicate the potential for errors or unforeseen circumstances during testing procedures. The selection of equipment, environmental conditions, calibration, observation, evaluation, analytical methods, testing protocols, and preparation can all mitigate these errors and uncertainties. It is essential to verify the dependability, probability and accuracy, of the experimental results. Consequently, an unforeseen analysis is conducted using the methodology outlined by Holman. The overall percentage of the experiment's uncertainty is calculated by taking the square root of the sum of the squares of the uncertainties for, HC, EGT BTE, CO, BSFC, NOx, BP and Pressure. Total experiment uncertainty = $\sqrt{}$ ((pressure transducer uncertainty)^2 + (angle encoder uncertainty)^2 + (NOx uncertainty)^2 + (HC uncertainty)^2 + (CO uncertainty)^2 + (EGT uncertainty)^2 + (stop watch uncertainty)^2 + (manometer uncertainty)^2.

$$\text{Total percentage of uncertainties} = \sqrt{}\{(0.1)^2 + (0.2)^2 + (0.1)^2 + (1)^2 + (0.2)^2 + (1.0)^2 + (0.15)^2 + (1.0)^2\} = \pm 2.3\%$$

III. RESULT AND DISCUSSION

III.1 COMPARISON OF VARIOUS HYDROGEN FLOW RATES WITH BIODIESEL

This text discusses the performance and emission characteristics of B15 Karanja biodiesel, which was initially utilized as fuel in CI engines with hydrogen flow rates of 6, 9, 12, 15, and 18 lpm. Graphs of emissions for HC, CO, and NOx are plotted versus load, while efficiency and BSFC are utilized to create performance graphs.

III.2 EFFECT ON BRAKE THERMAL EFFICIENCY (BTE)

Figure 3 illustrates the correlation between BP and BTE for B15 as a function of hydrogen flow rate. The optimal brake thermal efficiency, in comparison to diesel fuel, is attained with hydrogen flow rates of 15 lpm and 18 lpm combined with B15. Consequently, it may be inferred that hydrogen actively participates in combustion and is responsible for enhancing thermal efficiency. The enhancement of brake thermal efficiency results from hydrogen's numerous advantageous properties, such as its elevated calorific value, swift combustion characteristics, and its ease of amalgamation with air. Blends of B15 with 15 lpm and 18 lpm exhibited improved brake thermal efficiency compared to diesel, at 28.60% and 27.62%, respectively.

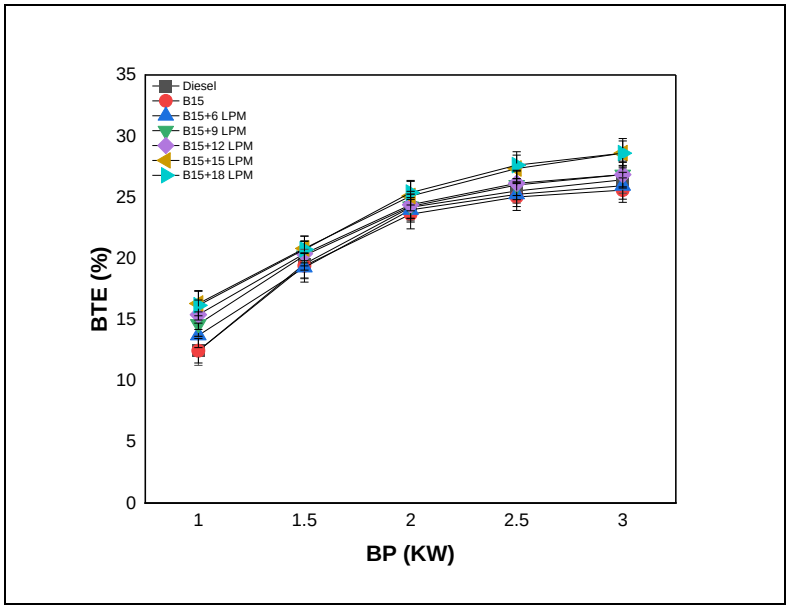


Figure 3: The B15 blend's BP and BTE at varying hydrogen mass flow rates.
Source: Authors, (2026).

In another study, A preheater was fabricated to raise intake air temperatures (30–60°C) for B20 Pongamia biodiesel. Higher intake air temperature improved BTE by 1% and reduced BSFC by 9%, while lowering HC by 36% and smoke by 16%. However, NOx emissions increased by 10.1%. [24].

III.3 EFFECT ON BSFC

Figure 4 depicts the correlation between BP and BSFC for the B15 blend at varying H2 flow rates. For B15 + 15 lpm and B15 + 18 lpm blends, it is seen that brake specific fuel consumption decreases with an increase in hydrogen flow rate. The minimum BSFC for B15 at 15 lpm is 0.23 kg/kW-hr, and for B15 at 18 lpm, it is also 0.20 kg/kW-hr. B15 BSFC demonstrates superior performance under maximum load circumstances compared to traditional fuel, with varying hydrogen flow rates. When hydrogen and air are properly combined, the flame velocity can be enhanced, resulting in more efficient combustion. The B15 blend surpasses other blends with a BSFC of 15 lpm and 18 lpm. The optimal mixture for all loads is B15, containing a hydrogen concentration of 15 liters per minute.

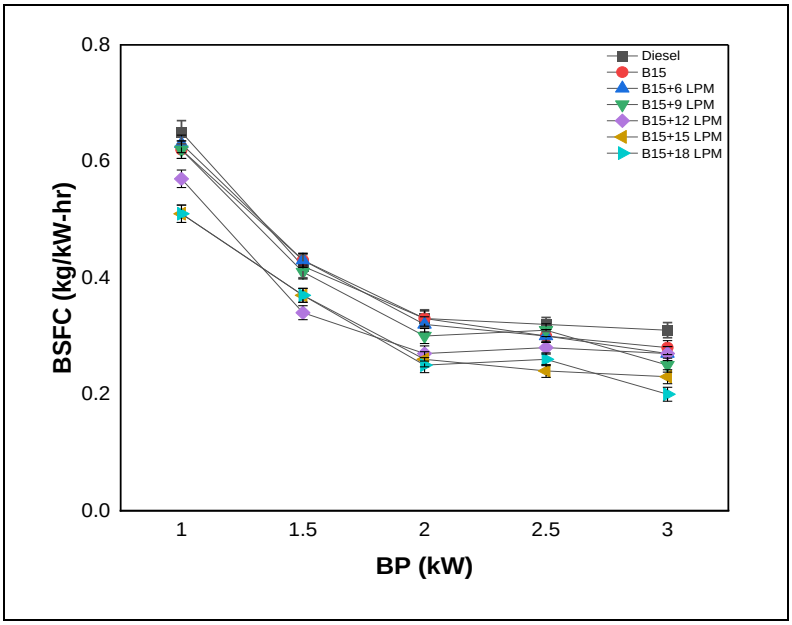


Figure 4: Comparison of BP and BSFC for the B15 blend at various hydrogen mass flow rates.
Source: Authors, (2026).

III.4 EFFECT ON HC

Figure 5 illustrates the hydrocarbon (HC) levels in relation to boiling point (BP) for B15 at various H2 flow rates. HC emissions were reduced in the B15 + 15 lpm and B15 + 18 lpm blends relative to standard diesel. The cylinder holds a blend of hydrogen and Karanja biodiesel oil, ensuring adequate mixing in the engine, elevating temperature, and facilitating the post-flame oxidation process. The hydrocarbon emission levels for diesel are 115 ppm, 75 ppm, and 80 ppm for B15 + 15 lpm and B15 + 18 lpm, respectively. B15 combined with 15 lpm decreases HC by 32.17%, whereas B15 combined with 18 lpm reduces it by 26.96%, respectively. In another study, Pongamia Pinnata methyl ester B20 was enhanced with 40–100 mg/L graphene oxide (GO) nanoparticles and 3% diethyl ether. At 100 mg/L GO + 5% ether, BTE rise by 19.7% and BSFC fell by 10.71% at maximum load. CO, HC, and smoke dropped by up to 70%, while combustion improved with higher peak pressure and shorter ignition delay [25].

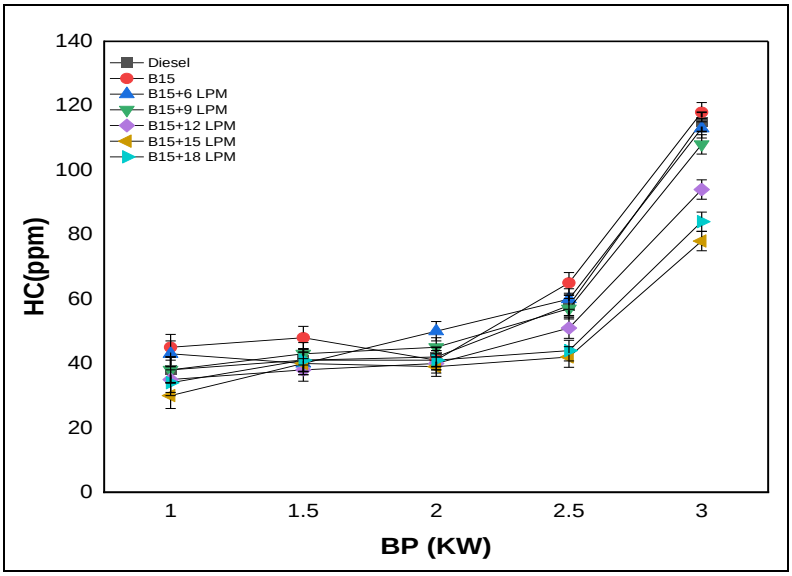


Figure 5: Variation of HC emissions with BP for the B15 blend at various hydrogen mass flow rates.
Source: Authors, (2026).

III.5 EFFECT ON CO

Variations in the H2 flow rate result in alterations to the CO/BP deviation for the B15 blend (refer to Figure. 6). Karanja biodiesel B15 lowers CO emissions with an increase in hydrogen flow rate. Reduced CO emissions were seen with the B15 + 15lpm and B15 + 18lpm blends in comparison to the others. This decreases CO emissions as the B15 + H2 mixture contains no carbon molecules. As a result, a substantial decrease in CO emissions is attained at all H2 flow rates. Diesel, B15 + 15 lpm, and B15 + 18 lpm exhibited CO emissions of 0.18%, 0.06%, and 0.17%, respectively. B15 combined with 15 lpm decreases CO emissions by 30% relative to diesel, while B15 with 18 lpm results in a 39% reduction. Prolonging combustion beyond the combustion stroke and subsequent extended injection results in a significant rise in CO levels at maximum load. As a result, when the exhaust valve was opened, the carbon monoxide emissions from the intermittent combustion products would solidify.

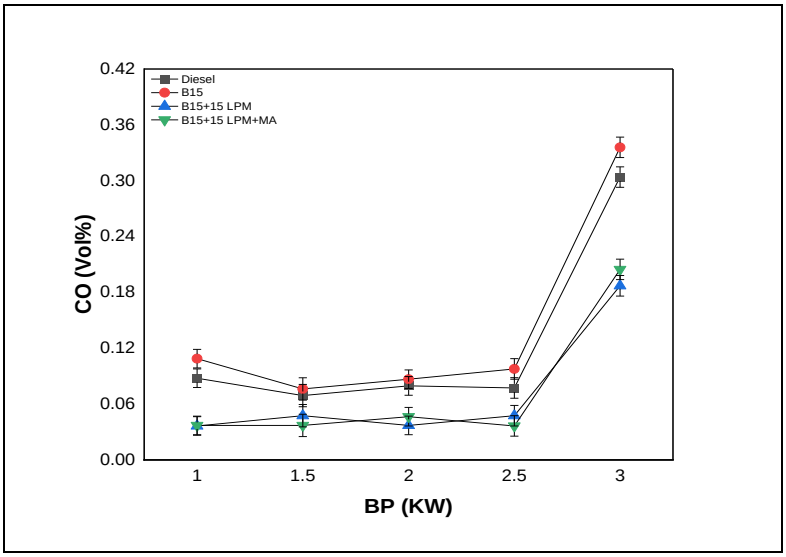


Figure 6: Variation of BSFC with BP the B15 blend at different hydrogen mass flow rates.
Source: Authors, (2026).

III.6 EFFECT ON NOX

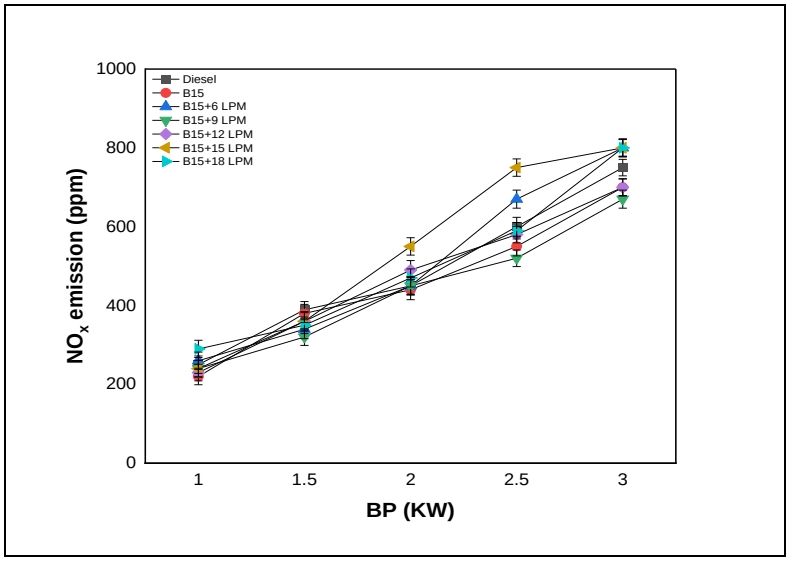


Figure 7: Emission of B15 blend with varying H2 flow rates as a function of BP and NOX.
Source: Authors, (2026).

Figure 7 illustrates the NOx levels in relation to BP for various H2 flow rates in the B15 mix. The complete response of an alternate fuel in the engine is governed by NOx, the most pivotal factor among these elements. As hydrogen flow rates increase, NOx levels also escalate. The production of nitrogen oxides (NOx) is primarily influenced by the temperature of the combustion chamber and the internal oxygen composition. Moreover, the presence of additional oxygen in the B15 blend exacerbates the production of NOx emissions. The combustion residence time is reduced due to the diesel engine's exceptionally narrow flame response zone. Diesel with B15 at 15 lpm exhibits NOx levels of 967.26 ppm, whereas B15 at 18 lpm shows values of 1307.72 ppm and 1340.58 ppm, respectively. The blends B15 + 15 lpm and B15 + 18 lpm exhibit increases of 25.2% and 27.6%, respectively, in comparison to diesel fuel. In earlier study, the use of TiO_2 nanoparticles (25–100 mg/L) in a B30 blend of Pongamia Pinnata biodiesel was studied. The B30T75 sample delivered the best performance. NOx emissions increased slightly (2.91%), and HRR rise by 5.4% compared to B30, indicating improved combustion efficiency [26].

III.7 INVESTIGATION ON NATURAL ANTIOXIDANT ADDITIVE

The optimal test fuel was identified as B15 + 15 lpm in the previous investigation. Nonetheless, irrespective of the load, this fuel generates a higher volume of NOx. Consequently, diesel, B15, and B15 + 15 lpm are evaluated using an antioxidant component inherently found in the B15 mixture. After using a natural additive, we can notice changes in performance and emissions. Following the same procedure as before, graphs of performance and emissions are produced.

III.8 EFFECT ON BTE

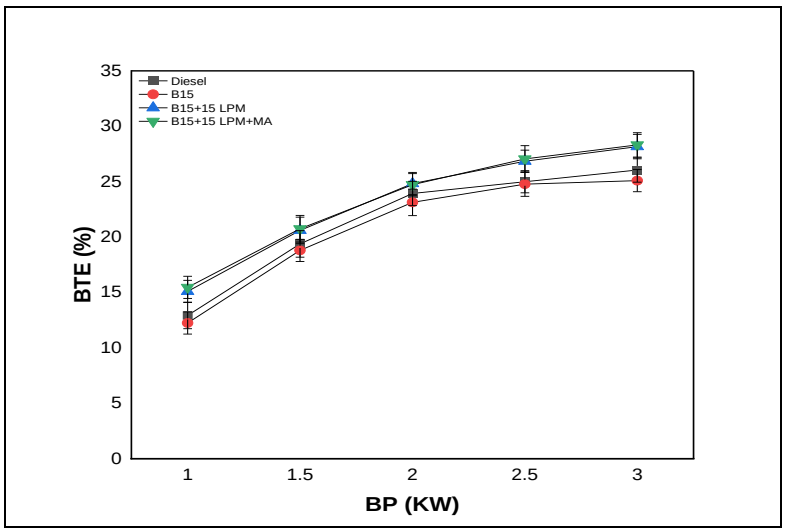


Figure 8: Comparison of B15 + 15 lpm BP against BTE with and without MA.
Source: Authors, (2026).

Both with and without the inclusion of a natural antioxidant component, Figure 8 shows BP with BTE for B15 and H2. The B15 + 15 lpm combination exhibits a Brake Thermal Efficiency (BTE) of 28.96%, which increases to 29.58% when an antioxidant ingredient (AI) is used. When operating at full load, the BTE efficiency of B15+ hydrogen with AI surpasses that of diesel by 5.2%. In previous study, Jojoba (JME), polanga (PME), and mixed methyl ester (MME) biodiesels tested enhanced with diglyme (cetane improver) and cumene (antioxidant) in a single-cylinder DI diesel engine at 1500 rpm. MME15+DG5 achieved the highest BTE improvement of 15.78%. Cumene also improved efficiency, with MME15+CU5 showing an 8.77% BTE increase and reduced exhaust gas temperature [27].

III.9 EFFECT ON BSFC

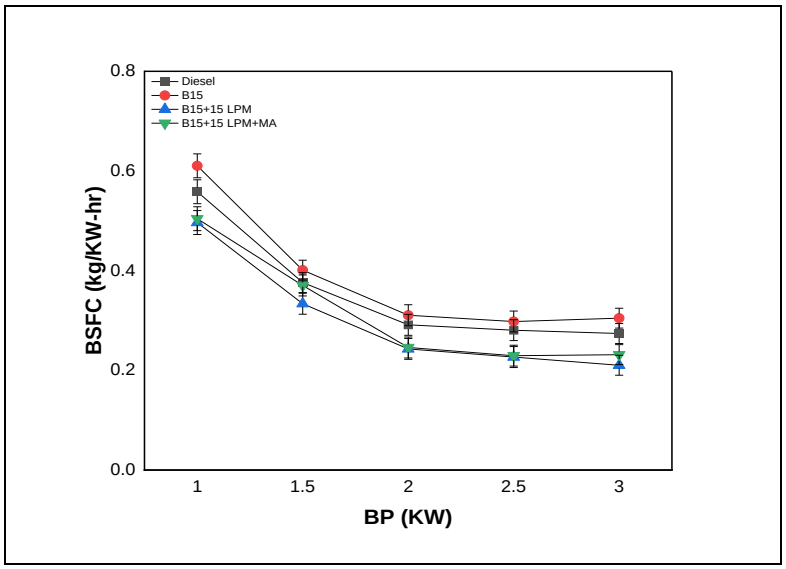


Figure 9 Comparison of BP and BSFC for B15 + 15 lpm with and without MA.
Source: Authors, (2026).

Both with and without the addition of the natural antioxidant MA, Figure 9 shows BP with BSFC for B15 and H2. Incorporating a natural antioxidant into a B15 blend with hydrogen in dual fuel mode in compression ignition engines enhances the brake specific fuel consumption (BSFC). The brake specific fuel consumption (BSFC) for B15 at 15 liters per minute with AI is 0.23152 kg/kW-hr, but without AI it is 0.21023 kg/kW-hr. The presence of chlorophyll and carotenoids results in a greater BSFC for B15 + 15 lpm compared to the additive blend.

In another study, Diphenylamine (DPA) antioxidant with ceria nanoparticles in a B30 Jatropha blend to address biodiesel's NOx emissions. Using hybrid RSM–MCDM (entropy, TOPSIS, VIKOR), the optimal B30+DPA50+$CeO_2$50 blend improved BTE by 1.01%, lowered BSFC by 5.67%, and reduced NOx by 18.87% [28].

III.10 EFFECT ON CO

Figure 10 illustrates the impact of CO on BP and B15 with H2, both with and without the inclusion of MA, a natural antioxidant. The addition of an antioxidant slightly increases CO levels. Incorporating 1000 ppm of additive into B15 + 15 lpm increases CO emissions from 0.187 08% to 0.43%. This happens because the natural antioxidant component takes too long to convert carbon monoxide to carbon dioxide and the combustion process is inadequate. The reduction in CO emissions can be attributed to the absence of carbon molecules in H2.

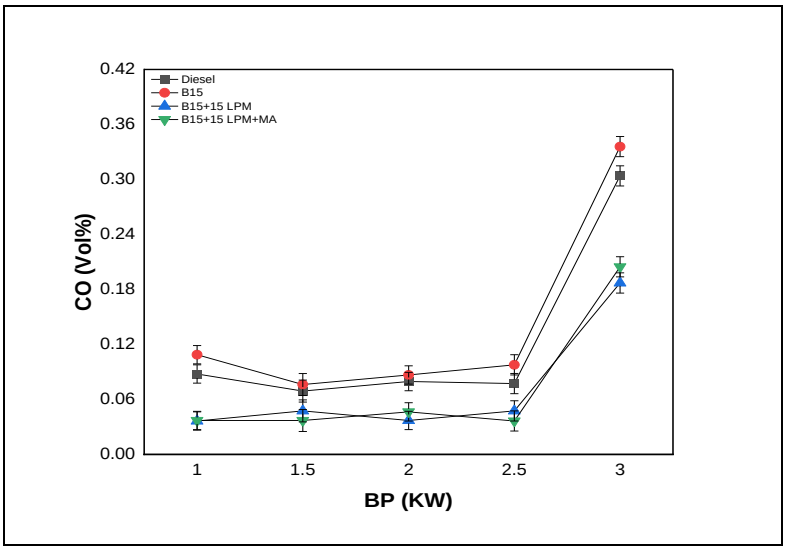


Figure 10: Comparison of B15 + 15 lpm's BP and CO with and without MA.
Source: Authors, (2026).

III.11 EFFECT ON HC

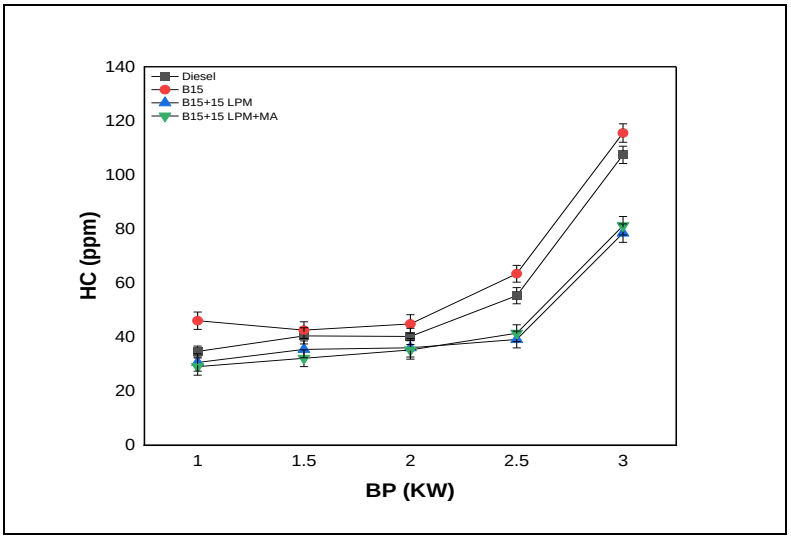


Figure 11: Comparison of B15 + 15 lpm BP and HC with and without MA.
Source: Authors, (2026).

Figure 11 illustrates the comparative effects of BP and HC on B15 and H2, both with and without the incorporation of MA, a natural antioxidant. A reduction in combustion chamber temperature is attributed to inadequate mixture formation, which leads to elevated HC emissions for B15 with 1000 ppm and 15 lpm of hydrogen. In comparing B15 + 15 lpm with and without the AI additive, the rise in HC for the former is 2.36 %. Furthermore, HC is measured at 79 ppm for the B15 + 15 lpm blend and 82 ppm for the B15 + 15 lpm blend with MA. In previous study, Albizia lebbeck leaf-derived natural antioxidant (NA) was incorporated into SBME20 (soybean methyl ester) biodiesel for low heat rejection (LHR) engines at 500–2000 ppm concentrations. The 1000 ppm blend delivered the best results, reducing BSFC to 0.38 kg/kWh and increasing BTE to 32%, alongside decreases in CO (−15%), HC (−10%), NOx (−12%), and smoke opacity (−20%) [29].

III.12 EFFECT ON NOX

Figure 12 illustrates both the NOx B15 and H2 BP samples, with and without the incorporation of the natural antioxidant agent MA. The NOx emissions from a B15 + 15 lpm blend are 24.14 % greater than those from conventional diesel, which is already significantly elevated. To reduce NOx emissions, add 1000 ppm of a natural ingredient to the B15 mixture. To reduce NOx emissions, the temperature and oxygen concentrations in the combustion chamber are diminished. The nitrogen oxides (NOx) for B15 + 15 lpm blends at 1000 ppm, with no load, are 1371.27 ppm and 1074.0 ppm, respectively. B15 combined with 15 lpm of additive decreases NOx emissions by 21.7 % relative to B15 with 15 lpm without the addition.

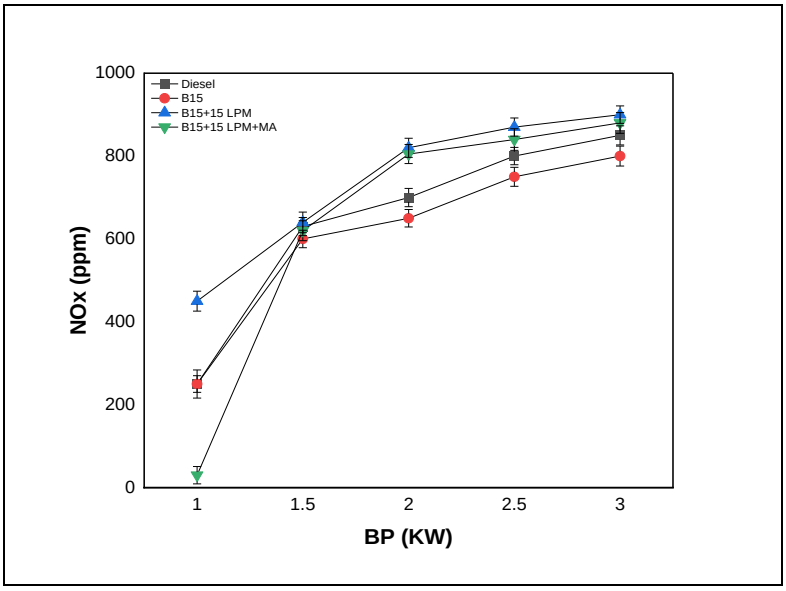


Figure 12: B15 + 15 lpm BP against NOx with and without MA.
Source: Authors, (2026).

IV. CONCLUSION

The subsequent results were obtained from tests performed on hydrogen and B15 in a CI engine, both with and without the use of 1000 ppm of AI additive.

- The B15 with 15 lpm and 18 lpm hydrogen shown an enhancement in brake thermal efficiency (BTE) and a reduction in brake specific fuel consumption (BSFC) compared to diesel. The BTE for B15 + 16 lpm, B15 + 18 lpm blends is 28.96 % and 29.58 %, respectively.
- In comparison to the B15 blend, the Brake Specific Fuel Consumption (BSFC) for the B15 + 15 lpm and B15 + 18 lpm mixtures is reduced to 0.23 kg/kW-hr and 0.20 kg/kW-hr, respectively.
- At maximum capacity, the CO emissions for the B15 + 15 lpm blends are diminished by 26.98%.
- HC emissions are diminished by 32% when B15 is amalgamated with 15 lpm of H2.
- As the flow rate of H2 increases, NOx emissions also rise.
- The outcomes for B15 at 15 lpm under maximum load are 1307 ppm, but the outcomes for B15 at 18 lpm are 1340 ppm. All measures enhanced when B15 and 15 lpm H2 with 1000 ppm MA were utilized.
- When B15 + 15 lpm with additive is compared to B15 + 15 lpm without additive, the BTE is elevated by 2.62 %. Both combinations exceed diesel in terms of brake thermal efficiency (BTE).
- Radioactive symbol B15 + 15 lpm without additive exhibits a BSFC of 0.2315 kg/Kw-hr, which is 8.8% higher than B15 + 15 lpm with additive.
- The B15 + H2 + MA blend exhibits a little increase in HC and CO emissions.
- The incorporation of antioxidant compounds leads to a substantial reduction in NOx emissions. In comparison to B15 + 15 lpm without additive, the mixture B15 + 15 lpm + AI decreases NOx emissions by up to 21.7%.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Kumaran P*, Natarajan S , Sathiyaraj S, Shanmugasundaram K and Subramanian T.
Methodology: Kumaran P*, Natarajan S , Sathiyaraj S, Shanmugasundaram K and Subramanian T.
Investigation: Kumaran P*, Natarajan S , Sathiyaraj S, Shanmugasundaram K and Subramanian T.
Discussion Kumaran P*, Natarajan S , Sathiyaraj S, Shanmugasundaram K and Subramanian T.
Writing: Kumaran P*, Natarajan S , Sathiyaraj S, Shanmugasundaram K and Subramanian T.
Writing: Kumaran P*, Natarajan S , Sathiyaraj S, Shanmugasundaram K and Subramanian T.
Resources: Kumaran P*, Natarajan S , Sathiyaraj S, Shanmugasundaram K and Subramanian T.
Supervision: Kumaran P*, Natarajan S , Sathiyaraj S, Shanmugasundaram K and Subramanian T.
Approval of the final text: Kumaran P*, Natarajan S , Sathiyaraj S, Shanmugasundaram K and Subramanian T.

VI. REFERENCES

[1] R. Dhairiyasamy and E. Murugesan, “Reducing NOx emissions in biodiesel-powered LHR engines using an antioxidant additive: a methodological approach,” International Journal of Ambient Energy, vol. 45, no. 1, 2024, doi: 10.1080/01430750.2023.2279139.

[2] P. Sivaraman, V. Kolandaivel, M. Natesan, and R. Dhairiyasamy, “Emission Reduction and Efficiency Gains in LHR Engines Using Natural Antioxidants in Biodiesel Blends,” International Journal of Automotive Technology, 2025, doi: 10.1007/s12239-025-00211-1.

[3] V. Kumar and A. K. Choudhary, “Prediction of the Performance and emission characteristics of diesel engine using diphenylamine Antioxidant and ceria nanoparticle additives with biodiesel based on machine learning,” Energy, vol. 301, 2024, doi: 10.1016/j.energy.2024.131746.

[4] R. Gowrishankar, P. M. Sithar Selvam, K. S. Narayana, T. Raja, R. Gokul, and S. Mayakannan, “Design and Optimization of a Biocompatible Polyethylene Composite with Nelumbo Nucifera Fiber and Nano-banana Peel Filler for Sustainable Environmental Applications,” Journal of Environmental Nanotechnology, vol. 14, no. 1, pp. 306–322, 2025, doi: 10.13074/jent.2025.03.24410.

[5] K. K. Pandey and M. Sivalingam, “Investigation of low heat rejection diesel engine run on N,N′-diphenyl-p-phenylenediamine antioxidant doped Jatropha methyl ester-diesel blend,” Environ Prog Sustain Energy, vol. 42, no. 3, 2023, doi: 10.1002/ep.14056.

[6] L. Razzaq et al., “Effect of biodiesel-dimethyl carbonate blends on engine performance, combustion and emission characteristics,” Alexandria Engineering Journal, vol. 61, no. 7, pp. 5111–5121, 2022, doi: 10.1016/j.aej.2021.10.015.

[7] M. Govindasamy, M. Ezhumalai, A. Munimathan, S. Dixit, S. Singh, and R. Dhairiyasamy, “Lemongrass oil as a renewable additive for enhancing the oxidation and thermal properties of calophyllum inophyllum biodiesel,” Results in Engineering, vol. 26, 2025, doi: 10.1016/j.rineng.2025.105102.

[8] S. R. Kandavalli et al., “Bioresin based hybrid green composite preparation using Holoptelea integrifolia fibers reinforced by Ziziphus jujuba seed particles: a fuzzy logic assisted optimization of mechanical behaviour,” Zeitschrift fur Physikalische Chemie, vol. 239, no. 5, pp. 747–772, 2025, doi: 10.1515/zpch-2024-0837.

[9] H. A. Dhahad and M. A. Fayad, “Role of different antioxidants additions to renewable fuels on NOX emissions reduction and smoke number in direct injection diesel engine,” Fuel, vol. 279, 2020, doi: 10.1016/j.fuel.2020.118384.

[10] A. Pugazhendhi et al., “Biodiesel from Scenedesmus species: Engine performance, emission characteristics, corrosion inhibition and bioanalysis,” Fuel, vol. 276, 2020, doi: 10.1016/j.fuel.2020.118074.

[11] F. Ahmed, B. Chetia, and S. Debbarma, “Effect of TBHQ antioxidant on the CRDI engine’s characteristics using hydrogen-enriched waste cooking biodiesel,” Int J Hydrogen Energy, vol. 160, 2025, doi: 10.1016/j.ijhydene.2025.150464.

[12] V. Kumar and A. K. Choudhary, “Influence of phenolic antioxidant additives on performance and emission characteristics of diesel engine fuelled with Jatropha biodiesel: A sustainable hybrid model using RSM and ANFIS,” Proc Inst Mech Eng C J Mech Eng Sci, vol. 238, no. 14, pp. 7253–7274, 2024, doi: 10.1177/09544062231224527.

[13] A. P. Edlabadkar, B. Mudadla, K. Karthick, M. S. Kumar, A. Adams, and M. Selvaraju, “Optimization of Machining Parameters and Performance Analysis of AA2024/ZrO2 Metal Matrix Composite Using TOPSIS: Insights into Squeeze Casting
and Tribological Behavior,” Journal of Environmental Nanotechnology, vol. 14, no. 1, pp. 181–192, 2025, doi: 10.13074/jent.2025.03.2441016.

[14] R. M. Reddy et al., “Experimental investigation on thermal barrier-coated diesel engine using antioxidant-doped orange peel oil biodiesel,” Surface Review and Letters, vol. 32, no. 9, 2025, doi: 10.1142/S0218625X24501270.

[15] S. Ramalingam, T. Rajkumar, S. Subramanian, and S. Palani, “Investigation of combustion, emission, and performance parameters of a natural antioxidant additives using hydrogen and biodiesel as dual fuel in CI engine operation,” Int J Hydrogen Energy, vol. 110, pp. 44–54, 2024, doi: 10.1016/j.ijhydene.2025.01.425.

[16] K. Yadav, N. Kumar, and R. Chaudhary, “Comprehensive Study of Additives and Corrosion in Biodiesel,” in Smart Innovation, Systems and Technologies, 2020, pp. 1–11. doi: 10.1007/978-981-15-2647-3_1.

[17] K. K. Pandey and M. S, “Effect of synthetic antioxidant-doped biodiesel in the low heat rejection engine,” Biofuels, vol. 14, no. 3, pp. 243–258, 2023, doi: 10.1080/17597269.2022.2132717.

[18] G. Uğuz, A. Çakmak, C. D. S. Bento, and N. Türköz Karakullukçu, “Experimental investigation of fuel properties and engine operation with natural and synthetic antioxidants added to biodiesel,” Biofuels, vol. 14, no. 4, pp. 405–420, 2023, doi: 10.1080/17597269.2022.2156049.

[19] J. R. Bikkavolu et al., “Predicting Common Rail Direct Injection (CRDI) engine metrics using nanoparticle-enhanced pongamia pinnata biodiesel with machine learning,” Emergent Mater, 2025, doi: 10.1007/s42247-025-01175-9.

[20] H. Narendrakumar Pandya, S. Banka, and S. Prakashbhai Parikh, “Pongamia pinnata biodiesel production using cobalt doped ZnO nanoparticles—An analytical study,” Environ Prog Sustain Energy, vol. 41, no. 4, 2022, doi: 10.1002/ep.13814.

[21] N. Jose, D. V Nair, K. Kavalli, R. Ramegowda, and G. S. Hebbar, “A study of diesel and Pongamia Pinnata biodiesel combustion in compression ignition engines using zero-dimensional modelling and experimental methods,” International Journal of Ambient Energy, vol. 43, no. 1, pp. 3335–3344, 2022, doi: 10.1080/01430750.2020.1831596.

[22] H. R. Amriya Tasneem, K. P. Ravikumar, H. V Ramakrishna, and B. Kuldeep, “Ceramic Material for Thermal Barrier Coatings in Compression Ignition Engine for its Performance Evaluation with Biodiesel,” in Materials Today: Proceedings, 2021, pp. 7745–7751. doi: 10.1016/j.matpr.2021.02.274.

[23] R. Dhairiyasamy, S. Rajukkannu, D. Varshney, S. Singh, D. Gabiriel, and E. Murugesan, “Examining the Effects of Neem Leaf Extract Additive on Biodiesel Blends for Improved Fuel Efficiency and Engine Performance in a Low Heat Rejection Engine,” Journal of Environmental Nanotechnology, vol. 14, no. 1, pp. 543–557, 2025, doi: 10.13074/jent.2025.03.2511294.

[24] R. Dsilva, B. K. Gopalakrishna, R. DSouza, V. Kaliveer, and T. Bhat, “Empirical study of engine performance and emission parameters using Python module,” in AIP Conference Proceedings, 2020. doi: 10.1063/5.0007072.

[25] M. E. M. Soudagar et al., “Utilization of non-edible bio-feedstock Pongamia Pinnata-diethyl ether ternary fuel blend supplemented with graphene oxide nanoparticles on CRDi engine characteristics,” J Therm Anal Calorim, vol. 149, no. 11, pp. 5687–5712, 2024, doi: 10.1007/s10973-024-13143-2.

[26] R. D. Silva, K. G. Binu, and T. Bhat, “Investigation of the Effect for a Higher Biodiesel Blend and Titanium Dioxide Nanoadditive on the Performance Parameters of CI Engine,” in AIP Conference Proceedings, 2024. doi:
10.1063/5.0196128.

[27] S. K. Nayak, D. B. Munuswamy, G. Subbiah, M. Naresh, and Y. Devarajan, “Influence of diglyme and cumene additives upon emission and combustion behaviour of diverse biodiesel fuelled diesel engine,” Results in Engineering, vol. 25, 2025, doi: 10.1016/j.rineng.2025.104255.

[28] V. Kumar and A. K. Choudhary, “A sustainable model using RSM and MCDM techniques to evaluate performance and emission characteristics of a diesel engine fueled with diphenylamine antioxidant and CeO2 nanoparticle additive biodiesel blends,” Journal of Renewable and Sustainable Energy, vol. 15, no. 6, 2023, doi: 10.1063/5.0168854.

[29] M. Balamurugan, R. Dhairiyasamy, W. Bunpheng, C. C. Kit, and D. Gabiriel, “Enhanced performance and reduced emissions in LHR engines using Albizia lebbeck antioxidant-infused SBME20 biodiesel,” Ind Crops Prod, vol. 222, 2024, doi: 10.1016/j.indcrop.2024.119677.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 976-983. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3143

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

DYNAMIC VIBRATION SUPPRESSION OF SMART COMPOSITE PLATES VIA LQR AND FINITE ELEMENT METHOD

Mourad Chalane*[1], Mohamed Latrache[2] and Bachir Labiodh[3]

[1]Department of Civil Engineering, Mohamed El-Bachir El-Ibrahimi University of Bordj Bou Arreridj, El-Anasser, 34030, Algeria
[2]Department of Mechanics, Faculty of Technology, University of M'sila, Algeria
[3]Department of Civil Engineering, Faculty of Technology, University of El Oued, Algeria

[1]http://orcid.org/0009-0002-1042-5654, [2]http://orcid.org/0009-0006-5438-2317, [3]https://orcid.org/0000-0003-0225-3290

Email: *mourad.chalane@univ-bba.dz, mohamed.latrache@univ-msila.dz, labiodh-bachir@univ-eloued.dz

ARTICLE INFO

Article History
Received: December 15, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Intelligent composite plate,
First-Order Shear Deformation Theory (FSDT),
Piezoelectric actuators and sensors,
Active Vibration Control (AVC),
Linear Quadratic Regulator (LQR).

ABSTRACT

In this paper, a comprehensive finite element (FE) formulation is developed for the dynamic analysis and active control of classical laminated composite plates with embedded piezoelectric layers. The model, based on the First-Order Shear Deformation Theory (FSDT) and Hamilton's principle, accurately captures transverse shear effects and electromechanical coupling. Stiffness, mass, and piezoelectric contributions of actuator and sensor layers are explicitly included, providing a fully consistent multi-field representation of the intelligent structure. To enhance vibration performance, piezoelectric actuators are employed within an active vibration control (AVC) framework. A Linear Quadratic Regulator (LQR) controller is designed using the independent mode space approach, enabling efficient computation of optimal feedback gains for dominant vibration modes while keeping the control model reduced-order and computationally tractable. Numerical simulations on various laminated configurations show excellent agreement with ANSYS benchmark results in terms of natural frequencies and mode shapes. The LQR-based AVC strategy is shown to significantly reduce vibration amplitudes, with peak responses decreased by up to 92% within a few seconds and the damping ratio increased more than fivefold. Frequency-response analysis reveals up to 50% attenuation of resonance peaks, highlighting the synergistic effect of state-feedback control and piezoelectric coupling in increasing effective damping. These results confirm the effectiveness of the proposed FE-LQR approach in suppressing vibrations, accelerating structural stabilization, and enhancing the dynamic resilience of smart composite plates.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The increasing demand for lightweight, high-strength composite materials with integrated sensing and actuation capabilities has significantly expanded their applications in aerospace, automotive, and civil engineering [1], [2]. Laminated and sandwich composite plates embedded with piezoelectric layers have attracted substantial attention due to their ability to suppress vibrations and reduce noise [3], [4], [5]. However, the inherent anisotropy and high flexibility of such structures make them prone to excessive vibrations, necessitating the development of advanced control strategies to ensure structural stability [6], [7]. Active vibration control (AVC) techniques employing piezoelectric materials as sensors and actuators provide rapid response and efficient electromechanical coupling [8], [9].

Among these techniques, the Linear Quadratic Regulator (LQR) has emerged as a widely used method due to its ability to optimize control effort while maintaining system stability, effectively balancing vibration suppression with energy consumption [10], [11], [12]. Recent studies have demonstrated the effectiveness of LQR control in reducing peak vibration amplitudes by over 90% and enhancing damping ratios by several times compared to uncontrolled systems [13], [14]. A robust numerical modeling framework is crucial for designing an effective controller that accurately captures material anisotropy, layer interactions, and complex laminated geometries. The finite element method (FEM) provides a versatile tool for such purposes, allowing incorporation of both mechanical and electromechanical contributions of piezoelectric layers [15], [16], [17]. Recent works have extended FEM formulations to account for viscoelastic damping, functionally graded material (FGM) cores, and optimal placement of piezoelectric patches for enhanced vibration control [18], [19], [20]. Comparisons between FEM-based simulations and commercial software (e.g., ANSYS) have consistently confirmed the high accuracy of these models in predicting natural frequencies, mode shapes, and time responses of smart composite plates [15], [17], [18]. In this study, a finite element-based modeling framework is combined with an LQR-based active vibration control strategy for laminated composite plates embedded with piezoelectric patches. Numerical simulations demonstrate substantial reduction in vibration amplitudes, attenuation of resonance peaks, and accelerated structural stabilization. These results confirm the effectiveness of the proposed approach and contribute to the ongoing development of smart composite structures with enhanced dynamic performance and reliable active control capabilities.

II. FINITE ELEMENT FORMULATION

In this section, a finite element (FE) formulation is developed to model the mechanical and electromechanical behavior of laminated composite plates with embedded piezoelectric layers. The formulation combines the First-Order Shear Deformation Theory (FSDT) for accurate kinematics with piezoelectric constitutive equations, allowing the evaluation of mechanical displacements, strains, and electromechanical coupling. The FE discretization procedure is then presented to assemble the global system matrices for numerical analysis.

II.1 MECHANICAL DISPLACEMENTS AND STRAINS

The First-Order Shear Deformation Theory (FSDT) accounts for transverse shear deformation, which is ignored in Classical Plate Theory, to describe mechanical displacements and strains. By including the transverse shear effect, FSDT provides more accurate predictions for moderately thick laminated composite plates, compared to the classical theory which assumes that normals to the mid-plane remain perpendicular after deformation [21], [22]. This theory is widely used in the analysis of composite and smart structures due to its balance between accuracy and computational efficiency [23]–[25].

II.2 PIEZOELECTRIC CONSTITUTIVE EQUATIONS

The piezoelectric effect can be expressed by four pairs of equations linking the mechanical tensors of deformation $\{\mathcal{E}\}$ and stress $[\sigma]$ to the vectors of the electric field $[\mathrm{E}]$ and the induction (the electric displacement) $\{\mathrm{D}\}$. It is common to present these equations in the following form[26]–[29]:

$$\begin{cases} \{\sigma\} = [\mathrm{Q}]\{\varepsilon\} - [\mathrm{e}]^{\mathrm{T}}\{\mathrm{E}\} \\ \{\mathrm{D}\} = [\mathrm{e}]\{\varepsilon\} - [\epsilon]\{\mathrm{E}\} \end{cases} \tag{1}$$

$[Q]$and $[e]$ are respectively called the elastic constants measured with constant electric field $\{E\}$, the piezoelectric constants and the measured dielectric constants with constant strain.

II.3 FINITE ELEMENT DISCRETIZATION

The shapes and structures of these matrices can be expressed in a standard coupled form, which is widely adopted for finite element modeling of piezoelectric laminated plates and smart structures [30] :

Matrix of masses

$$\begin{aligned} [\mathrm{M_e}] &= \int [\mathrm{N}]^{\mathrm{T}}[\bar{\mathrm{m}}][\mathrm{N}]\mathrm{dv} = [\mathrm{M_{ep}}, \mathrm{M_{ap}}, \mathrm{M_{sp}}] \\ &= \int_{-1}^{1} \int_{-1}^{1} [\mathrm{N}]^{\mathrm{T}}[\bar{\mathrm{m}}][\mathrm{N}]|J|\mathrm{d}\xi\mathrm{d}\eta \\ &= \sum_{\mathrm{i}=1}^{\mathrm{n_g}} \sum_{\mathrm{j}=1}^{\mathrm{n_g}} [\mathrm{N}]^{\mathrm{T}}([\bar{\mathrm{m}}_{\mathrm{p}}, \bar{\mathrm{m}}_{\mathrm{ap}}, \bar{\mathrm{m}}_{\mathrm{sp}}])[\mathrm{N}]|J|\mathrm{d}\xi\mathrm{d}\eta \\ [\bar{\mathrm{m}}] &= \sum_{\mathrm{k}=1}^{\mathrm{n}} \int_{\mathrm{z_k}}^{\mathrm{z_{k+1}}} \begin{bmatrix} \mathrm{I_1} & 0 & 0 & \mathrm{I_2} & 0 \\ 0 & \mathrm{I_1} & 0 & 0 & \mathrm{I_2} \\ 0 & 0 & \mathrm{I_1} & 0 & 0 \\ \mathrm{I_2} & 0 & 0 & \mathrm{I_3} & 0 \\ 0 & \mathrm{I_2} & 0 & 0 & \mathrm{I_3} \end{bmatrix} \mathrm{dz}, (\mathrm{I_1}, \mathrm{I_2}, \mathrm{I_3}) \\ &= \sum_{\mathrm{k}=1}^{\mathrm{n}} \int_{\mathrm{z_{k+1}}}^{\mathrm{z_k}} \rho(1, \mathrm{z}, \mathrm{z}^2)\mathrm{dz} \end{aligned} \tag{2}$$

Rigidity matrix

$$\begin{aligned} [K] &= \int_{v} [B]^{T}[\overline{D}][B]dv = [K_{ep}, K_{ap}, K_{sp}] \\ &= \int_{-1}^{1} \int_{-1}^{1} [B]^{T}[D][B]|J|d\xi d\eta \\ &= \sum_{i=1}^{n_g} \sum_{j=1}^{n_g} [B]^{T}([\overline{D}_p, \overline{D}_{ap}, \overline{D}_{sp}])[B]|J|d\xi d\eta \end{aligned} \tag{3}$$

Damping Matrix

$$[\mathrm{C}] = \alpha[\mathrm{M}] + \beta[\mathrm{K}] \tag{04}$$

Electrical stiffness matrix

$$\begin{aligned}[\mathrm{K_{me}}] &= \int_{\mathrm{A}} [\mathrm{B}]^{\mathrm{T}}[\mathrm{e}]\left[\mathrm{B}_{\varphi}\right]\mathrm{dA} \\ &= \int_{-1}^{1}\int_{-1}^{1} [\mathrm{B}]^{\mathrm{T}}[\mathrm{e}]\left[\mathrm{B}_{\varphi}\right] \\ &= \sum_{\mathrm{i}=1}^{\mathrm{ng}}\sum_{\mathrm{j}=1}^{\mathrm{ng}} [\mathrm{B}]^{\mathrm{T}}[\mathrm{e}][\mathrm{B}_{\varphi}]|J|\mathrm{d}\xi\mathrm{d}\eta\end{aligned} \tag{05}$$

Coupling matrix

$$\begin{aligned}[K_{ea}, K_{es}] &= \int_{A} [B_{\varphi}]^{T}[\epsilon_{a,s}][B_{\varphi} \\ &= \int_{-1}^{1}\int_{-1}^{1} [B]^{T}[e_{a,s}][B_{\varphi}]|J|d\xi d \\ &= \sum_{i=1}^{n_g}\sum_{j=1}^{n_g} [B]^{T}[e][B_{\varphi}]|J|d\xi d\eta\end{aligned} \tag{06}$$

Mechanical forces

$$\begin{aligned}[\mathrm{F_m}] &= \int_{\mathrm{A}} [\mathrm{N}]^{\mathrm{T}}[\mathrm{d}]\mathrm{dA} = \int_{-1}^{1}\int_{-1}^{1} [\mathrm{N}]^{\mathrm{T}}[\mathrm{d}]|J|\mathrm{d}\xi\mathrm{d}\eta \\ &= \sum_{\mathrm{i}=1}^{\mathrm{ng}}\sum_{\mathrm{j}=1}^{\mathrm{ng}} [\mathrm{N}]^{\mathrm{T}}[\mathrm{d}]|J|\mathrm{d}\xi\mathrm{d}\eta\end{aligned} \tag{07}$$

III. ACTIVE VIBRATION CONTROL (AVC)

Active vibration control aims to suppress the dynamic response of flexible structures by using real-time feedback from sensors and applying appropriate control forces through actuators [31], [32] . In smart composite plates with integrated piezoelectric layers, this process requires an accurate state-space representation of the system to design an efficient controller such as the Linear Quadratic Regulator (LQR) [33], [34]. In this framework, the matrices [A], [B], and [C] correspond respectively to the system matrix, the control input matrix, and the output matrix. These matrices describe the structural dynamics, the influence of the actuator inputs, and the measured outputs used for feedback. They are derived by assembling the finite element mass, stiffness, and electromechanical coupling matrices into the standard state-space form [35], [36]:

The system matrix

$$\mathrm{A} = \begin{bmatrix} 0 & \mathrm{I} \\ -\mathrm{diag}(\omega_{\mathrm{k}}^2) & -2\xi_{\mathrm{k}}\omega_{\mathrm{k}} \end{bmatrix} \tag{08}$$

The control input matrix

$$\mathrm{B} = \begin{bmatrix} 0 \\ \mu\Omega^{\mathrm{T}}\mathrm{K}^{(\mathrm{ac})}{}_{\mathrm{me}} \end{bmatrix} \tag{09}$$

The output matrix

$$\mathrm{C} = [\mathrm{K}^{(\mathrm{i})}{}_{\mathrm{me}}{}^{\mathrm{T}}\Omega \quad 0] \tag{10}$$

This state-space formulation provides the basis for implementing the LQR controller and achieving effective vibration suppression in intelligent composite structures.

III.1 OBJECTIVE FUNCTION

To design the Linear Quadratic Regulator (LQR) compensator, the control problem is formulated as the minimization of a quadratic performance index [37],[38]. This objective function represents a trade-off between the vibration energy of the structure and the control effort required by the actuators. The quadratic cost function is classically defined as:

The quadratic cost function

$$\mathrm{J} = \int_{0}^{\infty}(\{\mathrm{X}\}^{\mathrm{T}}[\mathrm{Q}]\{\mathrm{X}\} + \{\varphi\}^{\mathrm{T}}[\mathrm{R}]\{\varphi\})\,\mathrm{dt} = min \tag{11}$$

Where: Q is a positive semidefinite matrix and R is a positive matrix.
The optimal solution is:

The optimal solution

$$[\mathrm{G}] = [\mathrm{R}]^{\mathrm{T}}[\mathrm{B}]^{\mathrm{T}}[\mathrm{K}] \tag{12}$$

Where [K] satisfies the Riccati equation:

The Riccati equation

$$[\mathrm{A}]^{\mathrm{T}}[\mathrm{K}] + [\mathrm{K}][\mathrm{A}] - [\mathrm{K}][\mathrm{B}][\mathrm{R}]^{-1}[\mathrm{B}]^{\mathrm{T}}[\mathrm{K}] + [\mathrm{Q}] = 0 \tag{13}$$

III.2 LOAD LINEAR QUADRATIC REGULATOR (LQR) PROBLEM

The Linear Quadratic Regulator (LQR) is a classical optimal control method used to compute a state-feedback control law that minimizes a quadratic cost function while ensuring system stability. In MATLAB, the command lqr is used to calculate the optimal gain matrix G.

The optimal gain matrix G

$$\text{Syntaxe} : [G, K, e] = Lqr(A, B, Q, R) \tag{14}$$

Where e is the closed-loop eigenvalues.

IV. NUMERICAL ANALYSIS AND DISCUSSIONS

In this section, some numerical results are presented for symmetrically laminated square plate [45, - 45, 45, - 45], with embedded four identical piezoelectric actuators. The geometrical size of the intelligent composite plate is L*L*H=1*1*0.002(m), for the piezoelectric patches are lp*lp*hp= 0.1*0.1*0.0001(m) (Fig.1). The localizations for the piezoelectric actuators are (X1, Y1) = (0.2, 0.4), (X2, Y2) = (0.4, 06), (X3, Y3) = (0.8, 0.7), (X4, Y4) = (0.4, 0.9) (Figure.1).

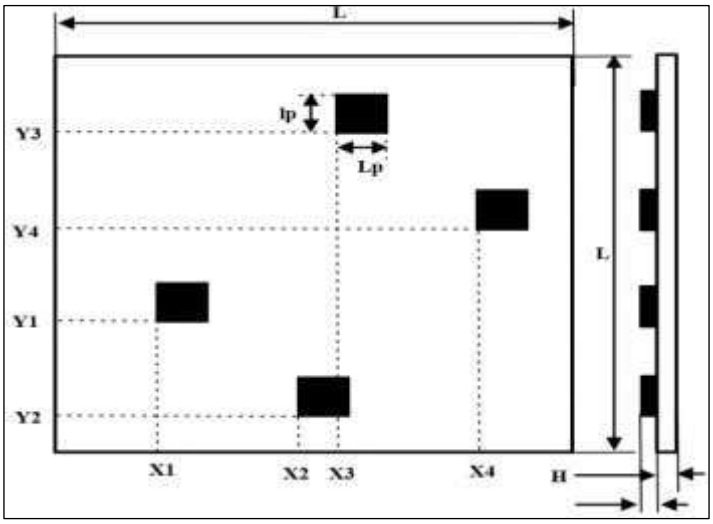


Figure 1: The geometrical model of the plate equipped with PZT patches.
Source: Authors, (2026).

The simulation featured in this paper assume α=0 and β=0.0015 damping constants. The time step Δt for Transient analysis is taken as $1 / (20 f_h)$, where f_h is the higher frequency. Consider an initial displacement field applied to the plate equal to 0.1 m. Table 1 contains the material properties data for the plate, piezoelectric actuators.

Table 1: Material properties.

properties	Graphite/epoxy	PZT G-1195
Poisson's ratio	0.31	0.3
Density ρ (kg/m3)	1550	7600
Elastic stiffness matrix (GPa) **E11** **E22** **E33** **G12** **G13** **G23**	 119 8.67 8.67 5.18 3.29 3.29	 132.38 10.76 10.76 3.61 5.61 5.61
Piezoelectric constant (C/m2) **e31** **e33** **e15**	-	 12.5 12.5 12.5
Dielectric constant(F/m) **g11** **g22** **g33**	-	 1.53×10-8 1.53×10-8 1.53×10-8

Source: Authors, (2025).

A finite element (FE) model, created using Matlab, facilitates the dynamic modeling of a system. Numerical results are verified using ANSYS, where an FE model of an intelligent composite plate is developed. The solid layers and PZT materials are modeled with SOLID45 and SOLID5 elements, featuring 8 nodes. A coupling electromechanical system is established using the CP command, and a suitable voltage potential is assigned (Figure 2).

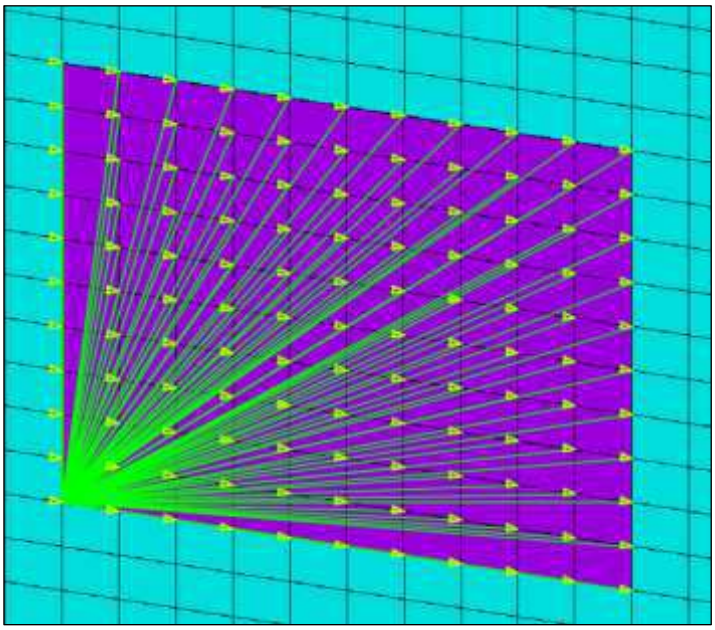


Figure 2: A coupling electromechanically created by CP command.
Source: Authors, (2026).

The current cantilever intelligent composite plate is meshed by global element size ebig=$10e^{-3}$, eight-node solid elements (Figure 3).

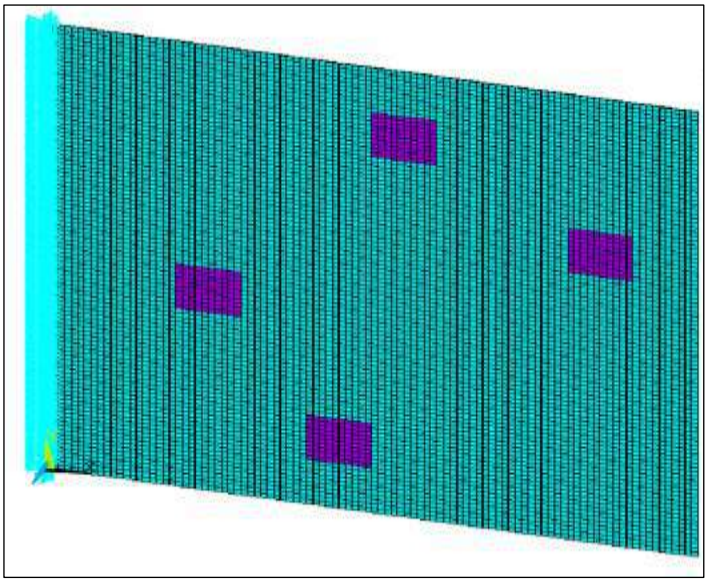


Figure 3: FEM of the cantilever intelligent composite plate and boundary conditions.
Source: Authors, (2026).

The modal analysis of the system reveals that the first three natural frequencies (Table 2) of the system lie in the range 110Hz.

Table 2: The first three natural frequencies (Hz) of the intelligent plate.

	Matlab	**Ansys**	**Error (%)**
First Mode	2.3219	2.5297	3.14
Second Mode	5.0515	5.7154	5.24
Third Mode	7.2065	7.3853	1.49

Source: Authors, (2026).

In this study, a linear quadratic optimal controller is considered to control the first three modes of the flexible plate. The weighting matrices chosen are: Q= $5e^{-8}$ and R=$0.5e^{+10}$. Here, the control is started after an elapse of 0.5 s in order to compare the controlled and uncontrolled responses. Figure 04 shows the displacement responses of the uncontrolled and controlled intelligent composite plate.

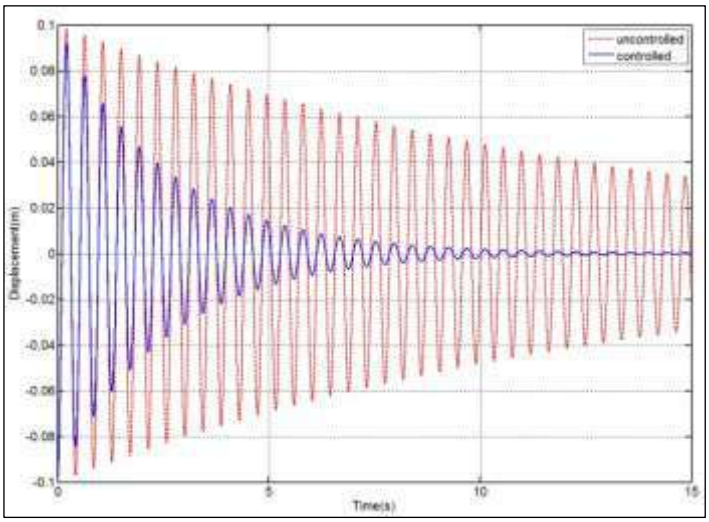


Figure 04: Comparison of uncontrolled and controlled displacement response of an intelligent composite plate with LQR controller.
Source: Authors, (2026).

The quantitative analysis of the responses shows that the introduction of LQR active control significantly improves the dynamic behavior of the composite plate. The maximum amplitude is reduced by approximately 36% from the initial instant, and by nearly 92% after 5 seconds, compared to the uncontrolled case. The calculation of the logarithmic decrement indicates that the damping ratio increases from about 0.61% (uncontrolled structure) to 3.3% under control, corresponding to an increase by a factor of ≈ 5.4. Moreover, the controlled system reaches a quasi-steady state in less than 4 seconds, whereas the uncontrolled system continues to vibrate with high amplitudes throughout the simulation window. These results clearly demonstrate the effectiveness of the LQR controller in enhancing effective damping, reducing vibrational energy, and accelerating the stabilization of the structure. The frequency response graph in Figure 5 compares the uncontrolled and controlled vibration amplitudes of the intelligent composite structure over the frequency range of 10–100 rad/s.

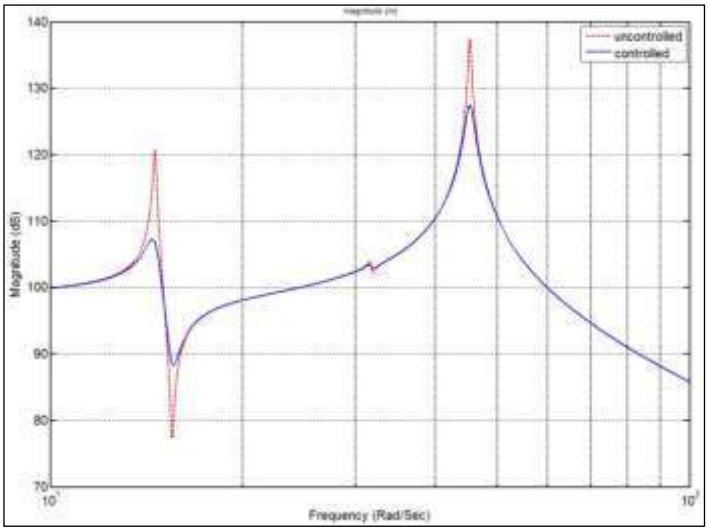


Figure 5: Comparison of the uncontrolled and controlled frequency response of an intelligent composite plate with LQR controller.
Source: Authors, (2026).

The controlled frequency-response curve exhibits a marked reduction in resonance amplification, with the main peak decreasing from approximately 133 dB (uncontrolled) to about 125 dB (controlled), corresponding to nearly a 6–8 dB attenuation which represents a reduction of about 40–50% in vibration energy; this flattening of the frequency response is consistent with the expected behavior of optimal state-feedback controllers, and is further enhanced by the piezoelectric coupling mechanism through which part of the induced strain energy is transformed into electrical energy, thereby increasing the effective damping of the structure.

V. CONCLUSIONS

In conclusion, the results demonstrate that the combined finite element modeling and LQR control strategy effectively reduces vibrations in smart composite plates with piezoelectric layers. Comparison of simulations in MATLAB and ANSYS confirms the accuracy of the model, as the natural frequencies match closely with minimal error. The numerical analyses further show that the LQR controller significantly enhances the dynamic performance of the plate, reducing peak vibration amplitudes by up to 92% within a few seconds and increasing the damping ratio more than fivefold. Frequency-response analysis highlights a substantial attenuation of resonance peaks, with up to 50% reduction in vibration energy, reflecting the synergistic effect of state-feedback control and piezoelectric coupling in increasing effective damping. Overall, these findings validate the LQR controller's effectiveness in suppressing vibrations, accelerating structural stabilization, and improving the dynamic resilience of smart composite structures.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Mourad Chalane*, Mohamed Latrache and Bachir Labiodh.
Methodology: Mourad Chalane*, Mohamed Latrache and Bachir Labiodh.
Investigation: Mourad Chalane*, Mohamed Latrache and Bachir Labiodh.
Discussion Mourad Chalane*, Mohamed Latrache and Bachir Labiodh.
Writing: Mourad Chalane*, Mohamed Latrache and Bachir Labiodh.
Writing: Mourad Chalane*, Mohamed Latrache and Bachir Labiodh.
Resources: Mourad Chalane*, Mohamed Latrache and Bachir Labiodh.
Supervision: Mourad Chalane*, Mohamed Latrache and Bachir Labiodh.
Approval of the final text: Mourad Chalane*, Mohamed Latrache and Bachir Labiodh.

VII. REFERENCES

[1] Z. Huang, Y. Mao, A. Dai, M. Han, X. Wang & F. Chu, "Active Vibration Control of Piezoelectric Sandwich Plates," Materials, vol. 15, no. 11, article 3907, 2022, doi:10.3390/ma15113907.

[2] N. Malha, R. Tiberkak, M. Ezzraimi & M. Essedik Lazar, "Active vibration control of functionally graded material plates integrated with piezoelectric layers using new Q9γ approach," Journal of Mechanics, vol. 40, pp. 93–109, 2024, doi:10.1093/jom/ufae005.

[3] Z. Huang, H. Peng, X. Wang & F. Chu, "Modeling and Vibration Control of Sandwich Composite Plates," Materials, vol. 16, no. 3, article 896, 2023, doi:10.3390/ma16030896.

[4] Richumon Salu, G. P.V., Silpa Caroline James, "Active Vibration Control of Laminated Composite Thin Plates Using a Smart Material," International Journal of Engineering Research & Technology (IJERT), vol. 11, no. 2, 2023.

[5] X. Shao, G. Wei & A. Li, "Active vibration control for three-dimensional braided composite beams based on piezoelectric sensor and actuator," Journal of Mechanics of Materials and Structures, vol. 18, no. 4, pp. 503–519, 2023.

[6] A. Alsahlani, A. Al-Khateeb, M. Ahmed & A. Eidan, "Active Vibration Control of Piezoelectric Composite Plates Using Gain Scheduling Method," RCMA / IIETA, 2024.

[7] Z. Huang, Y. Liang, Y. Cheng & X. Wang, "Finite element modeling and vibration control of composite beams with partially covered active constrained layer damping," Sound & Vibration, 2024, 59(1), article 1765, doi:10.59400/sv1765.

[8] Z. Huang, H. Peng, X. Wang & F. Chu, "Finite Element Modeling and Vibration Control of Plates with Active Constrained Layer Damping Treatment," Materials, vol. 16, no. 4, article 1652, 2023, doi:10.3390/ma16041652.

[9] W. Sun, J. Du, P. Qiao, B. Qi, S. Yan & H. Tan, "Smart Active Vibration Control System of a Rotary Structure Using Piezoelectric Materials," Sensors, vol. 22, no. 15, article 5691, 2022, doi:10.3390/s22155691.

[10] "An embedded piezoelectric actuator for active vibration control: Concept, modeling, simulation, and investigation," Chinese Journal of Aeronautics, vol. 38, no. 4, article 103345, 2025, doi:10.1016/j.cja.2024.103345.

[11] Z. Huang, H. Peng & F. Chu, "Finite element modeling and active vibration control of functionally graded plates with piezoelectric patches," Materials, vol. 15, no. 6, 2022.

[12] Z. Satla, L. Boumia & M. Kherrab, "Vibration control of FGM plate using optimally placed piezoelectric patches," Revista Mexicana de Física, vol. 70, article 011002, 2024, doi:10.31349/RevMexFis.70.011002.

[13] X. Li, J. Zhao & M. Wang, "Active vibration control of laminated composite plates using LQR and LQG methods," Journal of Sound and Vibration, 2023.

[14] Y. Chen, H. Liu & F. Zhang, "Piezoelectric-based active damping of composite sandwich plates under dynamic loading," Composite Structures, vol. 310, 2023.

[15] M. Khan, S. Ahmed, A. Hussain & R. Ali, "Finite element analysis of smart composite plates with piezoelectric sensors and actuators," Materials Today: Proceedings, 2022.

[16] H. Zhang, L. Wang & P. Sun, "Dynamic modeling and vibration control of laminated plates using FEM," Mechanical Systems and Signal Processing, 2023.

[17] L. Wang, X. Li & H. Zhang, "Comparison of FEM simulations and ANSYS results for piezoelectric smart plates," Journal of Vibration and Control, 2022.

[18] F. Liang, Y. Chen & X. Zhang, "Active control of sandwich plates with functionally graded cores and piezoelectric layers," Composite Structures, 2024.

[19] Y. Huang, X. Li & H. Liu, "Optimized placement of piezoelectric actuators for active vibration control of laminated plates," Smart Materials and Structures, 2023.

[20] J. Zhou, X. Wang & M. Li, "Finite element modeling and LQR control of piezoelectric smart composite structures," Materials, 2025.

[21] J. N. Reddy, Mechanics of Laminated Composite Plates and Shells: Theory and Analysis, 3rd Edition, CRC Press, 2020.

[22] P. V. Varadan, R. K. Batra, and A. Kumar, "Analysis of laminated composite plates using first-order shear deformation theory," Composite Structures, vol. 250, 113311, 2020, doi:10.1016/j.compstruct.2020.113311.

[23] S. R. Mallick and J. N. Reddy, "First-order shear deformation theory for thick composite laminates: Accuracy and applications," International Journal of Mechanical Sciences, vol. 193, 106114, 2021, doi:10.1016/j.ijmecsci.2021.106114.

[24] X. Li, Y. Huang, and Z. Zhang, "Dynamic analysis of laminated composite plates using first-order shear deformation theory," Composite Structures, vol. 256, 113168, 2021, doi:10.1016/j.compstruct.2020.113168.

[25] M. A. Khan, H. Zhang, and L. Wang, “Vibration and stress analysis of smart laminated plates with piezoelectric layers based on FSDT,” Mechanical Systems and Signal Processing, vol. 165, 108286, 2022, doi:10.1016/j.ymssp.2021.108286.

[26] IEEE Standards Association, “IEEE Standard on Piezoelectricity,” IEEE Std 176-1987, pp. 1–176, 1987, doi:10.1109/IEEESTD.1988.79638.

[27] T. Ikeda, Fundamentals of Piezoelectricity, Oxford University Press, Oxford, 1990.

[28] J. N. Reddy, “On the modeling of piezoelectric smart structures,” Journal of Intelligent Material Systems and Structures, vol. 10, no. 6, pp. 459–471, 1999, doi:10.1177/1045389X9901000603.

[29] N. W. Hagood and A. von Flotow, “Damping of structural vibrations with piezoelectric materials and passive electrical networks,” Journal of Sound and Vibration, vol. 146, no. 2, pp. 243–268, 1991, doi:10.1016/0022-460X(91)90525-H.

[30] Z. Huang, H. Peng, X. Wang, and F. Chu, “Finite element modeling and vibration control of composite plates with piezoelectric layers,” Materials, vol. 16, no. 4, article 1652, 2023, doi:10.3390/ma16041652.

[31] N. W. Hagood and A. von Flotow, “Damping of structural vibrations with piezoelectric materials and passive electrical networks,” Journal of Sound and Vibration, vol. 146, no. 2, pp. 243–268, 1991, doi:10.1016/0022-460X(91)90525-H.

[32] J. T. P. Yao, “Concept of structural control,” Journal of the Structural Division, ASCE, vol. 98, no. ST7, pp. 1567–1574, 1972.

[33] B. D. O. Anderson and J. B. Moore, Optimal Control: Linear Quadratic Methods, Prentice Hall, Englewood Cliffs, NJ, 1990.

[34] X. Li, Y. Chen, and Z. Wang, “Active vibration control of laminated composite plates using LQR and LQG methods,” Journal of Sound and Vibration, vol. 553, 117522, 2023, doi:10.1016/j.jsv.2023.117522.

[35] J. N. Reddy, “Finite element modeling of piezoelectric smart structures,” Journal of Intelligent Material Systems and Structures, vol. 12, no. 3, pp. 161–171, 2001, doi:10.1177/1045389X0101200301.

[36] Z. Huang, H. Peng, X. Wang, and F. Chu, “Finite element modeling and LQR-based active vibration control of piezoelectric laminated composite plates,” Materials, vol. 15, no. 6, article 2134, 2022, doi:10.3390/ma15062134.

[37] X. Li, Y. Chen, and Z. Wang, “Active vibration control of laminated composite plates using LQR and LQG methods,” Journal of Sound and Vibration, vol. 553, 117522, 2023, doi:10.1016/j.jsv.2023.117522.

[38] Z. Huang, H. Peng, and F. Chu, “Finite element modeling and LQR-based vibration control of piezoelectric laminated plates,” Materials, vol. 15, no. 6, article 2134, 2022, doi:10.3390/ma15062134.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 984-992. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3145

RESEARCH ARTICLE OPEN ACCESS

A LIGHTWEIGHT PRIVACY-PRESERVING DATA STORAGE USING MODIFIED RING SIGNATURE SCHEME BASED ON ELLIPTIC CURVE CRYPTOGRAPHY

Vasughi T.S*[1] and Muthulakshmi P[2]

[1, 2]Department of Computer Science, Faculty of Science and Humanities, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, India.

[1]http://orcid.org/0009-0000-5808-628X, [2]http://orcid.org/0000-0001-5422-6716

Email: *vt7068@srmist.edu.in, muthulap@srmist.edu.in

ARTICLE INFO

Article History
Received: December 15, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Anonymity
Blockchain
Elliptic curve cryptography
Ring signature
Unforgeability

ABSTRACT

Blockchain operates in a decentralized manner that provides the potential data available to the public. Privacy protection is an open challenge for data transparency. Applying Blockchain-based cryptographic algorithms in various applications can increase or protect data storage privacy. This paper constructs a lightweight privacy preserving data storage using Modified Ring signature Scheme based on Elliptic Curve Cryptography (MRSS-ECC). This solution built an anonymous user data storage to ensure the user information with highly secure and user identity privacy and validate the signature in the Blockchain network using the smart contract. Therefore, Security analysis of the scheme to verify correctness, unconditional anonymity, and unforgeability. The proposed system incorporates efficient key generation, signature formation, and verification processes to reduce computational overhead. Experimental evaluation demonstrates reduced signature size and improved verification time compared to related approaches. This framework can be adapted for e-governance, healthcare, and financial systems requiring secure, privacy-preserving public ledger integration.

Copyright ©2025 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

In the current blockchain technology, digital services such as electronic voting, electronic lottery, electronic bidding, electronic payment transactions of various virtual cryptocurrencies, and complaint reporting features are achieved by cryptographic construction such as ring signatures to facilitate protecting the anonymity of users. Chaum and van Heyst introduced the concept of group signatures, in which users in a group leak secrets; the signer's anonymity might weaken the anonymity of innocent users. Ring signature was first introduced by Rivest, Shamir, and Tauman [1] in 2001. [2] A ring signature is a specific cryptographic digital signature that simplifies the group signature. In the ring signature scheme, the signer randomly selects a user to conceal multiple numbers of members in the ring and no group managers with a unique key. At the same time, the signer combines the public key and private key to complete the signature without indicating their own identity. Only this signature set can be verified by the verifier, but it is not known who signed the signature. At present, the Blockchain network address of the data sender and receiver is a completely transparent transaction and raises the possibility of data privacy leaks; however, it ensures the authenticity of the on-chain data. Elliptic curve cryptography integrates with ring signature to provide high security and low computational load, plus it endows anonymity of the signers. The ring signature scheme will avoid the data privacy leaks of the data sender as well as the receiver can realize anonymous transactions on the blockchain. The proposed solution offers these security properties correctness, unconditional anonymity, and unforgeability. This paper proposes a lightweight privacy-preserving data storage based on the Ring signature scheme in Blockchain.

The ring signature is operating to secure the information of the transaction initiator and ensure the blockchain's privacy. The rest of the paper is structured as follows. Related works are presented in Section 2. Section 3,4 describes the proposed scheme as Modified Ring Signature Scheme based on Elliptic Curve Cryptography. Section 5 discusses the performance and evaluation metrics.

II. RELATED WORKS

Ring signatures are modern cryptographic primitives from the cryptographic community that are addressed as more significant attention for innovation and manufacturing societies due to the user privacy-preserving authentication feature and protection against malicious user actions. Ring signature also plays a vital role in many applications such as electronic payment, e-voting, and auctions. [3] proposed an e-voting system using a smart contract, threshold encryption, and linkable ring signature with a trusted bulletin board to provide greater privacy.[4]To preserve individual privacy when exchanging data combined with forward security and identity ring signature to increase data security. The regular ring signature remains undetectable double votes in the e-voting and anonymous transactions as they are unlinkable. A linkable ring signature solves this problem and ensures the linkability of two ring signatures signed by the same signer, but the identities of the signers remain anonymous. [5] Two valid signatures on the same message m with private or public key pair linked by the key image. A trusted third party detects the dishonest signer and provides double-spending protection for each transaction or voting. [6] A blockchain-based voting system uses a Paillier system to count the ballots without leaking the candidate's information. Proof of knowledge convinces the voting system that the vote cast by a voter is valid or not without revealing the content of the vote. A linkable ring signature ensures no one can trace the vote's owner and verify the genuine vote.[7] Efficient linkable and/or threshold ring signature without random oracles providing stronger anonymity and accountability. However, the signature size is sublinear with O (d.$\sqrt{n}$) where n is the number of users in the ring and d is the threshold value.

A blind signature is well-suited for creating untraceable systems in blockchain-enabled systems. A blind signature guarantees the anonymity of the user's sensitive information. The blind signature scheme has been in many applications as e-cash and e-voting, and untraceable payment systems, which are more suitable for privacy preservation in blockchain-enabled systems.[8] adopted an anti-quantum blind signature scheme based on the lattice assumption using reject sampling and the bimodal Gaussian distribution to improve security and efficiency and is more secure against quantum attacks.[9] Proposed a privacy-preserving scheme in a smart grid based on group blind signature to meet the requirements of Anonymous authentication and conditional anonymity without revealing the real identity and tracing the malicious users efficiently.[10]The signature size is independent of the ring size using a constant size ring signature scheme with an efficient implementation based on the strong RSA assumption in the random oracle model (ROM) for anonymous identification in ad-hoc groups.[11] Developed a brand-new supply chain traceability system that uses the MLSAG ring signature protocol to protect the privacy of sensitive data and achieves traceability by obscuring participants in publicly available data on previous transactions. The goal of using ring signature in privacy oriented cryptocurrencies is to protect the privacy of a sender of a transaction by making it computationally infeasible to determine the senders address given the signature [12]. The Verge currency employs stealth addressing, which enables a user to construct a one-time address based on a stealth address provided by the recipient to send money.

The address verified its ownership is only visible to the sender and recipient. Only the recipient has access to recover and use the funds at that particular address. [13] Another popular digital cryptocurrency Monero provides un-trackability means all possible signers have the same probability of each transaction. Tracing the sender transaction is protected to provide a higher level of privacy.[14] Dash is a digital cryptocurrency for protecting user privacy. It provides strong anonymity using a chained approach, PrivateSend mixing service reduces the correlation between addresses.[15] zcash anonymous cryptocurrency is needed to provide privacy, untraceability using zk-SNARKs. It provides a shield for the transaction addresses of the sender, receiver, and amount of coins sent. [16] Distributed key generation and elliptic curve cryptography are combined to provide ECC-based signature with stronger security. The use of distributed key agreement avoids to trust third party authorizer to distributed the key. The distributed generation center (KDC) is responsible for generating the system parameter, validate and manage the cluster members and verify the signature and upload them into the blockchain network.[17] A smart contract-based ring signature algorithm employs multiparty secure computation to preserver signer privacy. To prevent forgery "one encryption per signature is used and also address the issue of large signature size, reduced the signature size by two times and signature generation time and verification time by 3 three times compared to ECC ring signature[18]. A secure medical data sharing scheme is proposed that uses certificateless traceable ring signatures, IPFS for data storage, and smart contracts for access control.

Self-Controlling Objects (SCO) enable decryption and controlled data sharing, while distributed key generation ensures data integrity and preserves patient privacy. [19] An ECC-based ring signature is designed for blockchain to protect stored user data without revealing actual identities, only metadata is stored on-chain and smart contract enforces the privacy rules. This scheme is formally proven secure against unforgeability and anonymity under the random oracle model and also resists key expose attacks. This method also reduced signature size and computation time compared to pairing based schemes.[20] Linkable Ring Signature scheme with unconditional anonymity based on the e-NTRU problem under the random oracle model using preimage sampling, trapdoor generation and rejection sampling algorithm to improve the efficiency as well as shorter signature size. An aggregate ring signature is the combination of multiple ring signatures from different message and different rings into a short single signature to maintain the anonymity of each signer. It allows verification of single aggregated signature rather than of each signature of different ring, [21] smart grids face privacy risks from attacker, to address this issue, Certificate aggregate ring signature scheme provides unconditional anonymity, key management and remove key escrow, also ensure message integrity and efficiency with very low verification costs. [22] Two schemes are designed of privacy-focused blockchain transaction, first an aggregated ring signature that reduces signature space complexity with proven security. Second a compact multi message ring signature based on the lock principle to improve the efficiency. Aggregated linkable tags are proposed to detect double-spending in blockchain transactions.

II.1 PRELIMINARIES

Elliptic curve modulo a prime

Let q denote a large prime number, E denote an elliptic curve over an integer finite field F_q,the set of solutions $E_q(a, b) \in F_q$ defined an equation as follows:

$$y^2 = x^3+ax+b \ (mod\ q) \text{ and } 4a^3+27b^2 \# 0 \ (mod\ q) \text{ where a and b are constants } a, b, x, y \in F_q.$$

Suppose $P(x_1, y_1)$ and $Q(x_2, y_2) \in E_q(a,b)$. The point $P(x, y)$ on an elliptic curve that satisfies the $E_q(a, b)$. If $P = -Q$ the point $Q(x, -y)$ is the negative point of $P(x, y)$ The line l_n passes through P, Q and intersect the elliptic curve at a point $R'(x_3, y_3)$symmetrical about x axis then $P+Q = R(x_3, y_3)$ The point O called the point of infinity form an additive cyclic group

$$G=\{(x, y): a, b, x, y \in F_q, E(x, y)=0\} U \{O\}.$$

Definition 1:(Discrete logarithm problem). Let $G = (P) \leq g$. $P \in g$, P is a point with prime order q , Given a point $aP \in G$. Compute $a \in Zq^*$.

Security Model

The security requirements of a ring signature scheme, Consider the attacker **A** has access to the private keys of some users but also has access to all public keys of ring members.

Game1 Unforgeability of ring signature.

For Attacker A, $Succ_{RS,A}$ is defined as its probability of success in the following game between the challenger R and attacker A.

1. **Setup:** Given a security parameter l_n , challenger R runs the setup algorithm to obtain a list of system parameters and then challenger R sends system parameters to the attacker A.
2. **Hash Query:** Attacker A submits any value he chooses and challenger R returns the corresponding hash value to him.
3. **User Public Key Query:** The attacker A requests any user's public key whom he chooses and challenger R returns the corresponding public key pk_i to him.
4. **Private Key Query:** The attacker A requests a user's private key sk_i and the challenger R returns the corresponding private key sk_i .
5. **Ring Signature Query:** The attacker A submits any message he chooses, and challenger R returns the corresponding ring signature to him.
6. **Forge:** Eventually, the attacker outputs the ring signature σ* on a message m* such that

1. σ* is a valid signature
2. m* has never been submitted to the ring signature query

Definition 2:

A Forge $A(t, q_H, q_U, q_P, q_{RS}, \in)$ breaks a ring signature scheme means that if A runs in maximum time, A makes at most q_H hash queries, at most q_U users public key queries, at most q_P private key queries, at most q_{RS} ring signature queries then $Succ_{RS,A}$ is at least €. A ring signature scheme is $A(t, q_H, q_U, q_P, q_{RS}, \in)$ existentially unforgeable under an adaptively chosen message attack if no forger $A(t, q_H, q_U, q_P, q_{RS}, \in)$ breaks it.

Game 2 Anonymity of ring signature.

Let $U=U_1, U_2, \ldots\ldots U_n$ b be n signers. A is an attacker and R is Challenger who all are involved in game 2.

1. The challenger R runs the setup algorithm to obtain a list of system parameters and sends them to the attacker A.
2. The attacker A adaptively makes a polynomially bounded number of ring signature query.
3. In the challenge phase, the attacker outputs a message m, a group of n user's public key and two different public key $pk_1, pk_2 \in R$ to the challenger R. The challenger R randomly chooses a bit $\mu \in \{0,1\}$ and sends A to a ring signature $\sigma \leftarrow (m, R, x_\mu)$.
4. The attacker A can make a polynomially bounded number of ring signature queries.
5. Finally, Attacker A outputs a bit $\mu' \in \{0,1\}$.
6. Attacker A wins the above game if and only if $\mu=\mu'$.

Definition 3:

Define the probability of success in game 2 of attacker A Succ(A)= Pr[μ=μ′] = ½+∈. A ring signature is said to have unconditional anonymity if no attacker can win game 2 with a no non negligible probability advantage. This is to say, for A ring signature scheme is said to have unconditional anonymity if $\in=0$.

III. PROPOSED METHOD

This section constructs a MRSS-ECC for anonymous user data storage which can ensure the privacy of user information in the blockchain environment with high security and smart contract preset conditions is triggered to execute transactions T in the blockchain network. Furthermore, in order to increase user anonymity add the public key and private key random generation using Fisher-Yates shuffling algorithm during the signing messages. As illustrated in Algorithm 1, the specific model is described as follows:

1. **Setup:** Enter a security parameter l_n. l_n is a large prime number. Choose a large prime number $q_0 > l_n$ .The output parameter per={q_0, G_1, G_2, P, H_1, H_2, H_3} Where G_1 is a base point of elliptic curve E. G_2 is a large prime q_0 additive cyclic order group. P is the generator of G_2 Picks three secure hash functions for resisting collision H_1, H_2, H_3:

   $H_1 : E(F_q) \rightarrow E(F_q)$
   $H_2 : \{0,1\} \rightarrow F_q$
   $H_3 : \{0,1\}^* * G_1 \rightarrow Z_q^*$
   Z_q^* the finite field with q elements.

2. **KeyGen:** Let a users set as U_i $\{0,1\}^*$ bas random ring members $(U_1, U_2, \ldots\ldots U_n)$. blockchain system picks at random xZ_q^*band compute $pk_i \leftarrow x_i * P$. The user's public and private key (pk_i, sk_i).

3. **A ring:** The public key set $pk_i = pk_1, pk_2 . pk_n$. R is the set of ring members public key Randomly select $a, b, c \in Z_q^*$ and then calculate

$$L_i = \begin{cases} (a_i + b_i + c_i) * G & if\ i = s \\ (a_i + b_i + c_i) * G + (a_i + b_i + c_i) * pk_i & if\ i \neq s \end{cases} \quad (1)$$

$$R_i = \begin{cases} (a_i + b_i + c_i) * H_1(pk_i) & if\ i = s \\ (a_i + b_i + c_i) * H_1(pk_i) + (a_i + b_i + c_i) * KI & if\ i \neq s \end{cases} \quad (2)$$

Among them: $KI = sk_s * H_1 (pk_s)$, which serves as the message's signature picture and prevents double spending threats to the system. $H_1 (pk_i)$converts pk_i to a point on the finite field's elliptic curve. Randomly select $r \in Z_q^*$ and then calculate as follows

$$h = H_3(m \| r) \quad (3)$$

$$C = H_2(h, L_1, \ldots, L_n, R_1, \ldots, R_n) \quad (4)$$

The signer computes

$$\alpha_i = \begin{cases} C - \sum_{i=1}^{n} \alpha_i & if\ i = s \\ (a_i + b_i + c_i) * H_1(pk_i) + (a_i + b_i + c_i) * KI & if\ i \neq s \end{cases} \quad (5)$$

$$\beta_i = \begin{cases} (a_i + b_i + c_i) - \alpha_i * sk_i & if\ i = s \\ (a_i + b_i + c_i) & if\ i \neq s \end{cases} \quad (6)$$

4. **Averify :** Anyone of the verifiers can compute the transaction signature T_σ

$$\begin{cases} L_i' = \beta_i * G + \alpha_i * pk_i & i \in [1, n] \\ R_i' = \beta_i * H_1(pk_i) + \alpha_i * KI \end{cases} \quad (7)$$

Then check the signature is valid or invalid as follows:

$$\sum_{i=1}^{n} \alpha_i = H_2(h, L_i', \ldots, L_n', R_i', \ldots, R_n') \quad (8)$$

Algorithm 1: Modified Ring Signature Scheme based on Elliptic Curve Cryptography (MRSS-ECC).

Input: a security parameter
Output: open parameter per = (q_0, G_2, P, H_1, H_2, H_3);

1. Number of participants = n
2. **For** all n do
3. randomly shuffle the ring members
4. **End for**
5. **For** all U_i generate the public key and private key do
6. calculate $keyGen(par) \rightarrow (pk_i , sk_i)$
7. **End for**
8. signature image of the message to prevent double spend attack calculate
9. $KI = sk_s * H_1 (pk_s)$,
10. **For** each involved in the ring do
11. Randomly select $a, b, c \in Z_q^*$ and then generate a ring signature

Transaction initiator start signing and then calculate

Source: Authors, (2026).

$$L_i = \begin{cases} (a_i + b_i + c_i) * G & if\ i = s \\ (a_i + b_i + c_i) * G + (a_i + b_i + c_i) * pk_i & if\ i \neq s \end{cases}$$
$$R_i = \begin{cases} (a_i + b_i + c_i) * H_1(pk_i) & if\ i = s \\ (a_i + b_i + c_i) * H_1(pk_i) + (a_i + b_i + c_i) * KI & if i \neq s \end{cases}$$

Randomly select $r \in Z_q^*$ and then calculate as follows

$$h = H_3(m||r)$$
$$C = H_2(h, L_1, \ldots, L_n, R_1, \ldots, R_n)$$

The signer computes

$$\alpha_i = \begin{cases} C - \sum_{i=1}^{n} \alpha_i & if\ i = s \\ (a_i + b_i + c_i) * H_1(pk_i) + (a_i + b_i + c_i) * KI & if\ i \neq s \end{cases}$$
$$\beta_i = \begin{cases} (a_i + b_i + c_i) - \alpha_i * \mathrm{sk_i} & if\ i = s \\ (a_i + b_i + c_i) & if\ i \neq s \end{cases}$$

End for

12. Finally, the generated one-time ring signature becomes the transaction initiator s to the message m as $T_\sigma = (KI, \alpha_1, \alpha_2, \ldots \alpha_s \ldots . \alpha_n, \beta_1, \beta_2, \ldots \beta_s \ldots . \beta_n)$
13. **For** verify Anyone of the verifiers can compute the transaction signature T_σ do
14. **While**
15.

$$\begin{cases} L_i' = \beta_i * G + \alpha_i * pk_i & i \in [1, n] \\ R_i' = \beta_i * H_1(pk_i) + \alpha_i * KI \end{cases}$$

Then check the signature is valid or invalid as follows:

$$\sum_{i=1}^{n} \alpha_i = H_2(h, L_i', \ldots, L_n', R_i', \ldots, R_n')$$

16. **End while**
17. **End for**

Security Analysis

In this section, we mainly focus on the three aspects of correctness, unforgeability and unconditional anonymity.

Unforgeability Analysis

The ring signature Aring T_σ is generated by a signature algorithm. Since three hash function H_1, H_2, H_3 then $\alpha i = H2(h, L1, L2, \ldots \ldots Ln, R1, R2, \ldots \ldots Rn)$ and hash $h = H_3(m||r)$ are distributed over G and Z_q^* respectively. Randomly select $a, b, c \in Z_q^*$ and then calculate $L_i, R_i, \alpha_i, \beta_i$. Randomly select the private key of the signer $sk_i \in Z_q^*$. The probability of guessing the original signer is negligibly greater than 1/n. The attacker A only knows the public key and not others private key. The probability of guessing the original signer is negligibly greater than 1/n+1. So it is concluded that the attacker A has no advantage of guessing the original signer.

Theorem 1: If the discrete logarithm problem is hard, then the scheme is unforgettable against the attacker A in game 1 can adaptively choose a message to attack in the random oracle model.

Proof: Suppose the challenger R receives a random instance*(a,aP)* of the discrete logarithm problem and has to compute the value of a. Challenger R set the public key $pk_i^* = a*P$. R will run A as a subroutine and act as challenger of A in game 1. Without loss of generality, we assume that all the queries are distinct. Now we will show challenger R answer to the attacker A.

1. **Setup:** Given a security parameter l_n, challenger R runs the setup algorithm to obtain a list of system parameters and then challenger R sends system parameters to the attacker A.
2. **Hash Query:** Challenger R maintains the list $L(\alpha_i, \beta_i)$. The list is initially empty. When the attacker A makes a query $H(\alpha_i)$, challenger R selects a value β_i randomly and sets $H_1(\alpha_i) = \beta_i$. Then challenger R adds (α_i, β_i) to the h list and returns β_i to A.
3. **User Public Key Query:** Challenger R maintains the list $L(\alpha_i, \beta_i)$. The list is initially empty. When the attacker A makes a query $H(\alpha_i)$, challenger R selects a value β_i randomly and sets $H(\alpha_i) = \beta_i$. Then challenger R adds (α_i, β_i) to the L list and returns β_i to A.
4. **Private Key Query:** The attacker A makes a user's public key query if $pk_s \neq pk_i^*$, challenger R fails and stops. Otherwise challenger R returns sk_i to A.
5. **Ring Signature Query:** The attacker A submits any message m and a set of n users, and challenger R outputs T_σ ring signature. If there exists a user identity pk_s such that $pk_s \neq pk_i^*$, then the challenger R returns the ring signature T_σ by calling the signing algorithm, where the original signer. Otherwise, challenger R does as follows:

$$L_i = \begin{cases} (a_i + b_i + c_i) * G & if\ i = s \\ (a_i + b_i + c_i) * G + (a_i + b_i + c_i) * pk_i^* & if\ i \neq s \end{cases} \quad (9)$$

$$R_i = \begin{cases} (a_i + b_i + c_i) * H_1(pk_i) & if\ i = s \\ (a_i + b_i + c_i) * H_1(pk_i) + (a_i + b_i + c_i) * KI^* & if\ i \neq s \end{cases} \quad (10)$$

$$h = H_3(m||r) \quad (11)$$

$$C = H_2(h, L_1, \ldots, L_n, R_1, \ldots, R_n) \quad (12)$$

$$\alpha_i = \begin{cases} C - \sum_{i=1}^{n} \alpha_i & if\ i = s \\ (a_i + b_i + c_i) * H_1(pk_i) + (a_i + b_i + c_i) * KI^* & if\ i \neq s \end{cases} \quad (13)$$

$$\beta_i = \begin{cases} (a_i + b_i + c_i) - \alpha_i * sk_i^* & if\ i = s \\ (a_i + b_i + c_i) & if\ i \neq s \end{cases} \quad (14)$$

Finally the ring signature T_σ for message m

$$T_\sigma^* = (KI^*, \alpha_1^*, \alpha_2^*, \ldots \alpha_s^* \ldots \alpha_n^*, \beta_1^*, \beta_2^*, \ldots \beta_s^* \ldots \beta_n^*)$$

Unconditional Anonymity Analysis

Theorem 2: Our ring signature scheme has the property of unconditional anonymity of the signer, For any algorithm A,any set of signers $U = U_1, U_2, \ldots U_n$ and a random $U_s \in U$,the probability $Pr[\mu = \mu'] = ½$ where $T_\sigma = (KI, \alpha_1, \alpha_2, \ldots \alpha_s \ldots \alpha_n, \beta_1, \beta_2, \ldots \beta_s \ldots \beta_n)$is a ring signature U generated by U_s .

Proof:

1. The challenger R runs the setup algorithm to obtain a list of system parameters and sends them to the attacker A.
2. The attacker A adaptively makes a polynomial bounded number of ring signature query.
3. In the challenge phase, the attacker outputs a message m, a group of n user's public key and two different public keys $pk_1, pk_2 \in R$ to the challenger R. The challenger R randomly chooses a bit $\mu \in \{0,1\}$ and sends A to a ring signature $\sigma \leftarrow Aring(m, R, x_\mu)$.
4. The attacker A can make a polynomial bounded number of ring signature queries.
5. Finally, Attacker A outputs a bit $\mu' \in \{0,1\}$.
6. Attacker A wins the above game if and only if $\mu = \mu'$.

Correctness Analysis

Anyone of the verifiers can compute the transaction signature T_σ **if** it is true to this formula

$$\sum_{i=1}^{n} \alpha_i = H_2(h, L_i', \ldots, L_n', R_i', \ldots, R_n')$$

When $i = s$

$$\begin{gathered} L_i' = \beta_i * G + \alpha_i * pk_i \\ = [(a_i + b_i + c_i) - \alpha_i * \mathrm{sk_i}] * \mathrm{G} + \alpha_i * pk_i] \\ = [(a_i + b_i + c_i) * G - \alpha_i * \mathrm{sk_i} * \mathrm{G} + \alpha_i * pk_i] \\ = L_i \end{gathered} \quad (15)$$

$$\begin{gathered} R_i' = \beta_i * H_1(pk_i) + \alpha_i * KI \\ = [(a_i + b_i + c_i) - \alpha_i * \mathrm{sk_i}] * H_1(pk_i) + \alpha_i * \mathrm{sk_s} * H_1(pk_s) \\ = [(a_i + b_i + c_i) * H_1(pk_i) - \alpha i * \mathrm{sk_i} * H_1(pk_i) + \alpha_i * \mathrm{sk_s} * H_1(pk_s) \\ = [(ai + bi + ci) * H_1(pk_i)] \end{gathered}$$

$$= R_i \quad (16)$$

When i ≠ s

$$\begin{gathered} L_i' = \beta_i * G + \alpha_i * pk_i \\ = (a_i + b_i + c_i) * G + (a_i + b_i + c_i) * pk_i \\ = L_i \end{gathered} \quad (17)$$

$$Ri' = \beta i * \mathrm{H1}(\mathrm{pki}) + \alpha i * \mathrm{KI} = (a_i + b_i + c_i) * H_1(pk_i) - \alpha_i * \mathrm{KI} = R_i \quad (18)$$

$\mathrm{H_2}(\mathrm{h}, \mathrm{L_1}, \ldots, \mathrm{L_n}, \mathrm{R_1}, \ldots, \mathrm{R_n}) = H_2(h, L_i', \ldots, L_n', R_i', \ldots, R_n')$

$$= C_s + \sum_{i=1 i \neq s}^{n} \alpha_i$$
$$= \sum_{i=1}^{n} \alpha_i$$

IV. PERFORMANCE AND EVALUATION

This section discusses all experiments were conducted on a system equipped with an Intel Core @2.6GHz, 16GB RAM, running Ubuntu 22.04 LTS. The computational complexity of the proposed solution written in Python and the complexity of the most significant phases: key generation time, signing time, and verification time. The implementation used secp256k1 parameters for the elliptic curve scalar multiplication of group G over a 256-bit prime field. The cryptographic hash functions H1, H2, H3 were implemented using SHA-256 with H1 configured for hash-to-curve mapping. The computation running time of the number of ring members in each ring in milliseconds. The Ring size(n) of 5,10,20,50,100 of participants for plaintext message "Hello" were tested and average timings were recorded. The figure.1 shows the output of the signature validity. The figure.2 shows the number of ring members increases the corresponding signature time and verification of the ring signature is gradually increase that is shown in the linear relationship. The result shows that the verification of the ring signature consumes less time than the signature time.



Figure 1: Output of Signature Validity.
Source: Authors, (2026).

Table 1: Cryptography Operation Notations.

Notation in Time	Cryptography operation
T_{SME}	ECC-based Scalar Multiplication Operation
T_{AE}	ECC-based Point Addition Operation
T_{HP}	Map-to Point Operation
T_M	Multiplication Operation
T_P	Bilinear-pair Operation
T_E	Exponential Calculation Time

Source: Authors, (2026).

Table 2: Cryptography Operation time in milliseconds.

Cryptography Operation	T_{SME}	T_{AE}	T_{HP}	T_M	T_P	T_E
Execution Time(ms)	1.674	1.234	3.812	-----	-----	-----

Source: Authors, (2026).

In our proposed scheme, the user signature requires the ECC-based scalar multiplication operation 3+(n-1)*4=(4n-1)T_{SME} , ECC-based point Addition operation 2(n-1)T_{AE} ,a hash operation mapping to points on elliptic curves 1T_{HP} , the signature verification requires the ECC-based scalar multiplication operation 4nT_{SME} , ECC-based point Addition operation 2nT_{AE} ,a hash operation mapping to points on elliptic curves 1T_{HP} , and a one-way hash operation mapping to a finite field of prime numbers, with the last three having negligible computational overhead. Therefore, the computational overhead of generating and verification of a signature is

$$T_{sign} = (4n-1)\, T_{SME} + 2(n-1)T_{AE} + T_{HP}$$
$$T_{verify} = 4nT_{SME} + 2nT_{AE} + T_{HP}$$

Table 3: Efficiency analysis of Ring Signature scheme from Elliptic Curve Group.

Algorithm	Signature Generation	Signature Verification
[18]	$(4n-1)T_M + (4n+6)\, T_E$	$nT_E + 2T_P$
[16]	$4(n-1)\, T_{SME} + 2(n-1)\, T_{AE} + T_{HP} + T_M$	$4nT_{SME}$
[17]	$4(n-1)\, T_{SME} + 2(n-1)\, T_{AE} + 2nT_{HP}$	$4nT_{SME} + 2nT_{AE} + T_{HP} + T_M$
[OURS]	$4(n-1)\, T_{SME} + 2(n-1)\, T_{AE} + T_{HP}$	$4nT_{SME} + 2nT_{AE} + T_{HP}$

Source: Authors, (2026).

To conclude, the time consumption of different ring signature schemes in the signature-generation and signature-verification steps is summarized in Table 3. It was observed that our scheme took less time than the other schemes [16][17] [18], in the signature generation phase Compare with[18] [16] our scheme takes less signature time and slightly more verification time. This indicates that our signature scheme possesses higher signing and verification efficiency.

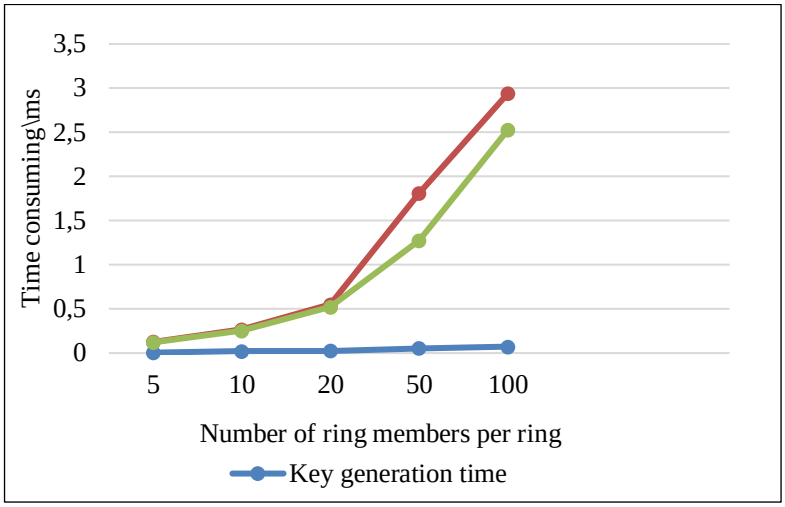


Figure 2: Number of Ring Members per ring for Ring signature scheme from Elliptic curve group.
Source: Authors, (2026).

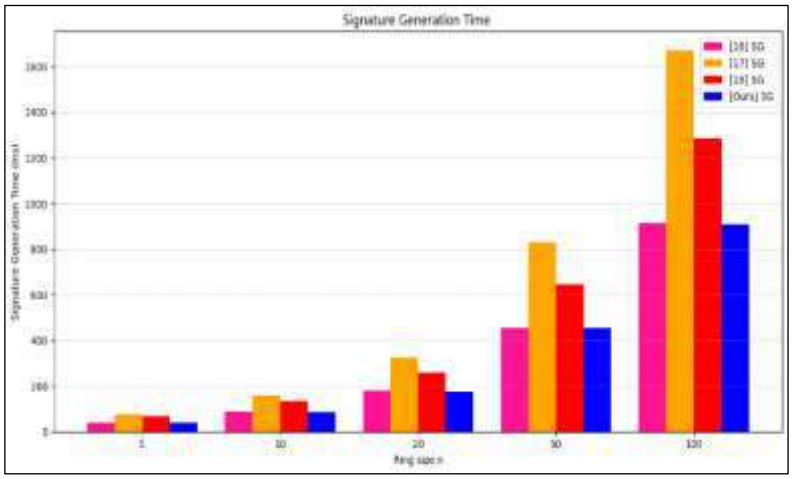


Figure 3: Comparison of Signature Generation time.
Source: Authors, (2026).

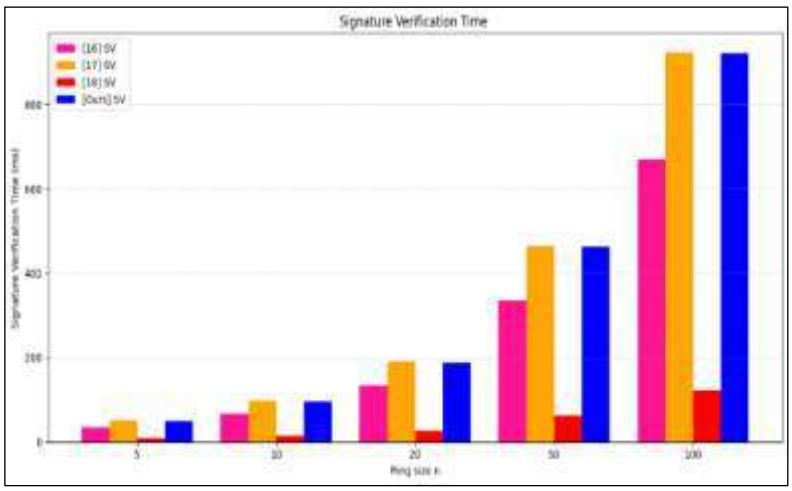


Figure 4: Comparison of Signature verification time.
Source: Authors, (2026).

IV. CONCLUSION

This article constructs a Modified Ring Signature Scheme based on the Elliptic Curve Cryptography (MRSS-ECC) for anonymous user data storage which can ensure the privacy of user information in the blockchain environment with high security and smart contract preset conditions is triggered to execute transactions T in the blockchain network. Our scheme achieves a concealment of user identity and provides anonymity, uniqueness, and unforgeability and can be applied in application with double spending protection. Finally, by comparing with related signature schemes, our scheme achieves faster signature generation and verification time and higher efficiency. However, since the efficiency and communication overhead grow as the number of users increases, future work will focus on developing an aggregated ring signature approach to enhance both verification speed and communication performance.

V. REFERENCES

[1] R. L. Rivest, A. Shamir, and Y. Tauman, “How to leak a secret,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 2248, pp. 552–565, 2001, doi: 10.1007/3-540-45682-1_32.

[2] M. N. S. Perera, T. Nakamura, M. Hashimoto, H. Yokoyama, C. M. Cheng, and K. Sakurai, “A Survey on Group Signatures and Ring Signatures: Traceability vs. Anonymity,” Cryptography, vol. 6, no. 1, pp. 1–22, 2022, doi: 10.3390/cryptography6010003.

[3] J. Lyu, Z. L. Jiang, X. Wang, Z. Nong, M. H. Au, and J. Fang, “A Secure Decentralized Trustless E-Voting System Based on Smart Contract,” 2019 18th IEEE Int. Conf. Trust. Secur. Priv. Comput. Commun. IEEE Int. Conf. Big Data Sci. Eng., pp. 570–577, 2019, doi: 10.1109/TrustCom/BigDataSE.2019.00082.

[4] J. Herranz, “Identity-based ring signatures from RSA,” vol. 389, pp. 100–117, 2007, doi: 10.1016/j.tcs.2007.08.002.

[5] L. Malina, J. Hajny, P. Dzurenda, and S. Ricci, “Lightweight ring signatures for decentralized privacy-preserving transactions,” ICETE 2018 - Proc. 15th Int. Jt. Conf. E-bus. Telecommun., vol. 2, no. Icete, pp. 526–531, 2018, doi: 10.5220/0006890505260531.

[6] B. Yu et al., “Platform-Independent Secure Blockchain-Based Voting System,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11060 LNCS, pp. 369–386, 2018, doi: 10.1007/978-3-319-99136-8_20.

[7] T. H. Yuen, J. K. Liu, M. H. Au, W. Susilo, and J. Zhou, “Efficient Linkable and/or Threshold Ring Signature Without Random Oracles,” Comput. J., vol. 56, no. 4, pp. 407–421, 2013, doi: 10.1093/comjnl/bxs115.

[8] C. Li, Y. Tian, X. Chen, and J. Li, “An efficient anti-quantum lattice-based blind signature for blockchain-enabled systems,” Inf. Sci. (Ny)., vol. 546, pp. 253–264, 2021, doi: 10.1016/j.ins.2020.08.032.

[9] J. P. D. Comput, W. Kong, J. Shen, P. Vijayakumar, Y. Cho, and V. Chang, “A practical group blind signature scheme for privacy protection in smart grid,” J. Parallel Distrib. Comput., vol. 136, pp. 29–39, 2020, doi: 10.1016/j.jpdc.2019.09.016.

[10] Q. Wu, W. Susilo, Y. Mu, and F. Zhang, “LNCS 4266 - Ad Hoc Group Signatures,” no. 60403007, pp. 120–135, 2006.

[11] M. El Maouchi, “DECOUPLES : A Decentralized , Unlinkable and Privacy-preserving Traceability System for the Supply Chain,” pp. 364–373, 2019.

[12] W. Koerhuis, T. Kechadi, and N. Le-khac, “Forensic Science International : Digital Investigation Forensic analysis of privacy-oriented cryptocurrencies,” Forensic Sci. Int. Digit. Investig., no. xxxx, p. 200891, 2019, doi: 10.1016/j.fsidi.2019.200891.

[13] S. Tople and P. Saxena, “A Traceability Analysis of Monero ’ s Blockchain,” pp. 153–173, 2017, doi: 10.1007/978-3-319-66399-9.

[14] E. Duffield and D. Diaz, “Whitepaper Dash : A Privacy-Centric Crypto-Currency”.

[15] A. Banerjee and M. Clear, “Demystifying the Role of zk-SNARKs in Zcash,” pp. 12–19, 2020.

[16] L. Wang, C. Peng, and W. Tan, “Secure Ring Signature Scheme for Privacy-Preserving blockchain,” Entropy, vol. 25, no. 9, p. 1334, Sep. 2023, doi: 10.3390/e25091334.

[17] Li, Q., Yi, W., Zhao, X., Yin, H., & Gerasimov, I. (2022). Representative Ring Signature Algorithm Based on Smart Contract. Sensors, 22(18). https://doi.org/10.3390/s22186805

[18] Lai, C.; Ma, Z.; Guo, R.; Zheng, D. Secure medical data sharing scheme based on traceable ring signature and blockchain. Peer Netw. Appl. 2022, 15, 1562–1576

[19] Li, X., Mei, Y., Gong, J., Xiang, F., & Sun, Z. (2020). A blockchain privacy protection scheme based on ring signature. IEEE Access, 8, 76765–76772. https://doi.or./10.1109/ACCESS.2020.2987831.

[20] Ye, Q., Wang, M., Meng, H., Xia, F., & Yan, X. (2022). Efficient Linkable Ring Signature Scheme over NTRU Lattice with Unconditional Anonymity. Computational Intelligence and Neuroscience, 2022. https://doi.org/10.1155/2022/8431874.

[21] Wang, H., Wang, L., Wen, M., Chen, K., & Luo, Y. (n.d.). A Lightweight Certificateless Aggregate Ring Signature Scheme for Privacy-preserving in Smart Grids.

[22] Teng, D., Yao, Y., & Huang, C. (2025). Optimizing signature space performance in privacy-enhanced blockchains: novel ring signature solutions. Eurasip Journal on Information Security, 2025(1). https://doi.org/10.1186/s13635-025-00192-9

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 993-1008. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3154

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

ENHANCING DISEASE DIAGNOSIS IN HEALTHCARE USING LSTM NETWORKS WITH OPTIMIZED FEATURE SELECTION: A COMPARATIVE RESEARCH ON HEART DISEASE, BREAST CANCER, AND LIVER DISEASE DATASETS

Tina Sachdeva*[1], Priyanka Sharma[2], and Dr. Mehtab Alam[3]

[1]Shaheed Rajguru College of Applied Sciences for Women, University of Delhi, Vasundhara Enclave, Delhi, India.
[2, 3]Acharya Narendra Dev College, University of Delhi, Govind Puri, Kalkaji, India

[1]http://orcid.org/0009-0003-8798-8719, [2]http://orcid.org/0009-0009-3341-6823, [3]http://orcid.org/0000-0001-7554-2160

Email: *tina.sachdeva@rajguru.du.ac.in, priyankasharma@andc.du.ac.in, mehtabalam@andc.du.ac.in

ARTICLE INFO

Article History
Received: December 17, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
LSTM model,
feature selection,
medical diagnosis,
heart disease,
breast cancer,
liver disorder

ABSTRACT

Clinical diagnosis relies heavily on accurate and timely medical assessment. However, clinical data sets are typically large, incomplete, and noisy, therefore limiting the reliability of traditional models used to diagnose clinical conditions. To address these limitations, this study developed an integrated model combining an optimised feature selection strategy with a Long Short-Term Memory (LSTM) model to provide improved accuracy across three clinical condition benchmarks, namely heart disease, breast cancer, and liver disorder. The first step in this process was feature selection, which involved selecting only those features that were deemed clinically relevant. By doing this, the subsequent LSTM model was able to recognise temporal patterns within the patient's clinical data set more easily than it would have been able to do using the original clinical data set. Results from this work demonstrates that the proposed methodology resulted in higher accuracy compared to baseline classifiers and also to non-optimised LSTM models across all three data sets. Specifically, the proposed method had an accuracy of 91.5% for heart disease, 97.6% for breast cancer, with a Receiver Operating Characteristic Area Under the Curve (ROC-AUC) of 0.99, and 83.9% accuracy for liver disorder. Additionally, there was a clear improvement in terms of recall and F1-score. Overall, the results from this study demonstrate that the integration of a method for dimensionality reduction with a sequential learning approach produces a reliable and generalizable clinical diagnostic model that can support clinical decision-making in various health care settings.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The increasing reliance on Machine Learning (ML) and Artificial Intelligence (AI) has made them essential components of contemporary health care, especially regarding automated and enhanced early disease detection [1]. Timeliness of diagnosis is also a very critical factor: an untreated cardiovascular condition may result in sudden death, delayed breast cancer detection leads to a lower survival rate, and untreated liver abnormalities will almost always result in irreparable damage. Manual analysis of laboratory and imaging data is the traditional basis for the majority of the diagnoses in clinical practice. While these approaches are successful, they suffer from human variability, subjectivity, and scalability issues in dealing with large numbers of patient data sets. Cardiovascular disease, breast cancer, and chronic liver disease are among the top reasons for morbidity and mortality globally. Cardiovascular disease is referred to as a "silent killer" because patients typically present with symptoms only after extensive physical damage [2]. The relationship between early detection and survival rates is well established for breast cancer, the most common form of cancer among women [3]. In the context of advanced liver disease, particularly in patients with underlying hepatitis or cirrhosis, complex portal vein thrombosis may arise and require specialised surgical reconstruction.[4].

Analysing these conditions as part of a unifying analytical framework helps to identify whether a single intelligent model can be generalised to different pathological areas. One of the persistent challenges in clinical analytics is the presence of poor or redundant variable information. Examples of features that are highly discriminatory in heart disease include cholesterol levels, blood pressure, and electrocardiogram (ECG) signals, whereas many peripheral features may have minimal effect on predicting outcomes. Maintaining irrelevant data contributes to the problem of the curse of dimensionality, causing overfitting and degrading the model's ability to generalise [5].

Techniques for feature selection can help alleviate the effects of redundancy by identifying the most influential predictor variables, reducing computational costs while providing clinicians with greater clarity of the relevance of retained variables. There is another aspect that is often neglected in clinical analytics, and that is the nature of the temporal evolution of the disease. Most diseases occur gradually, and isolated static measurements may hide underlying trends. For example, rising liver enzyme levels over time may signal early liver disease even though individual readings remain within normal limits. Likewise, subtle changes in ECG patterns observed over hospital visits may indicate impending cardiovascular events. To analyse the relationships in the sequences of clinical data, this research uses LSTM networks [6], a type of Recurrent Neural Network that retains long-range contextual information about sequences via special types of gates. Figure 1 depicts the overall structure of an LSTM.

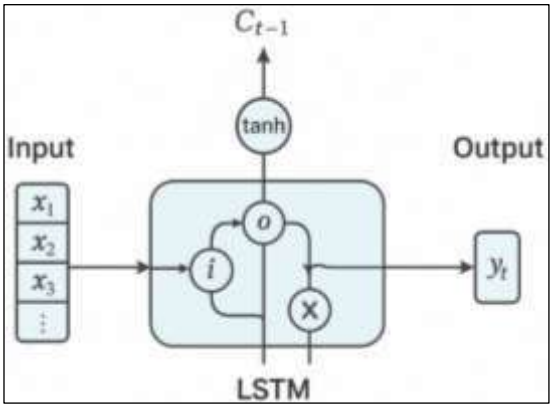


Figure 1: LSTM Network.
Source: Authors, (2026).

The work presented in this paper distinguishes itself from existing studies by introducing a tailored feature-selection approach followed by an LSTM model, and by evaluating their combined impact to enhance prediction performance across three publicly available datasets:

- **Heart Disease Dataset** (UCI Repository) [7]
- **Breast Cancer Dataset** (Wisconsin Diagnostic Dataset) [8]
- **Liver Disorder Dataset** (Indian Liver Patient Dataset) [9]

Maintaining consistency in the methodology used to evaluate each dataset allows this study to assess how optimisations to the feature space interact with temporal sequence modeling to improve clinical predictions. Maintaining a consistent evaluation methodology across all datasets enables this study to systematically examine how feature space optimisation interacts with temporal sequence modeling, ultimately contributing to improved clinical prediction performance.

I.1 RESEARCH QUESTIONS

Additionally, the following key questions guide this research:

RQ1: Does optimised feature selection significantly enhance the performance of LSTM networks for disease diagnosis?

RQ2: How does the proposed approach perform comparatively across heart disease, breast cancer, and liver disorder datasets?

RQ3: Which evaluation metric (accuracy, precision, recall, F1-score, ROC-AUC) shows the greatest improvement after feature optimisation?

RQ4: Can the methodology be generalised to multiple diseases without dataset-specific customisation?

I.2 CONTRIBUTIONS OF THIS RESEARCH

This research makes the following contributions:

- Proposes an **integrated framework** combining optimised feature selection with LSTM networks for medical diagnosis.
- Provides a **comparative evaluation** across three widely used benchmark datasets (heart disease, breast cancer, liver disease), highlighting generalizability.
- Demonstrates the role of **feature selection** in reducing computational complexity while improving classification accuracy.
- Offers **insights for healthcare practitioners** and data scientists on applying advanced deep learning methods in multi-disease contexts.

Table 1: Global Burden of the Targeted Diseases.

Disease	Global Prevalence (approx.)	Mortality Impact	Key Challenge in Diagnosis
Heart Disease [7]	523 million cases (2021)	~18 million deaths annually	Often asymptomatic until late stages
Breast Cancer [8]	2.3 million cases (2020)	~685,000 deaths annually	Requires early detection for survival
Liver Disease [9]	1.5 billion cases (2020, chronic liver conditions)	~2 million deaths annually	Complex diagnosis due to multiple causes

Source: Authors, (2026).

II. LITERATURE REVIEW

II.1 OVERVIEW AND SCOPE

This literature review focuses on two intersecting streams relevant to the proposed : (1) **deep sequence models** (particularly LSTM and hybrid LSTM architectures) applied to medical diagnosis (heart disease, breast cancer, liver disease and allied tasks), and (2) **feature selection and optimization** methods (filter, wrapper, hybrid, and metaheuristic optimizers such as Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) used to improve medical classification performance. Wherever available, preference is given to SCIE journals and high-quality open repositories (JAMA Network Open, Scientific Reports, BMC series, Expert Systems with Applications, Information Sciences, Sensors, etc.).

II.2 KEY THEMES FROM RECENT LITERATURE

II.2.1 LSTM and Hybrid LSTM Models for Medical Diagnosis

A growing number of studies deploy LSTM or hybrid Convolutional Neural Network (CNN–LSTM)/Conv1D–LSTM architectures for detection/classification tasks:

- Hybrid image-processing + sequence architectures (CNN + LSTM) show strong performance on breast cancer imaging and histopathology datasets [10]. Recent Scientific Reports and BMC Medical Imaging papers demonstrate near-state-of-the-art accuracy using CNN-LSTM hybrids.
- For cardiovascular or heart disease tasks, hybrid CNN–LSTM have been proposed to capture temporal patterns or sequential clinical features, with clear performance gains over shallow models [11].
- For liver disease and cirrhosis prognosis, deep sequence models have been used to model longitudinal clinical records and outperformed simple linear models in some cohort studies. JAMA Network Open and PLoS One showcase LSTM/RNN success in longitudinal hepatology prediction tasks [12].

LSTM (and LSTM hybrids) are widely validated across modalities — imaging (when combined with CNN front-ends), waveform data (ECG), and longitudinal clinical records — and routinely outperform classical linear/ensemble baselines when sequential information is present.

II.2.2 Feature Selection (Classical and Optimised) in Medical Diagnosis

Feature selection remains a central preprocessing step in medical ML systems:

- Metaheuristic methods such as PSO and GA remain popular for wrapper-style feature selection in clinical tasks; Sensors (2021) and several recent ESWA / Information Sciences papers detail PSO/GA variants tailored to feature selection with improvements in accuracy and dimensionality reduction [13].
- Hybrid and multi-objective approaches that balance predictive performance and subset size (e.g., multi-objective GA, hybrid filter-wrapper pipelines) show improved robustness across datasets. Recent Applied Soft Computing / ScienceDirect works present dynamic and hybrid feature selection schemes for medical diagnosis [14].
- Newer filter+grouped/structured feature selectors and niche GA variants address high-dimensional gene expression/imaging features for cancer diagnostics, indicating that domain-aware selection (grouping correlated features) improves interpretability and stability [15].

Metaheuristic and hybrid selection methods are effective in producing compact, high-quality feature sets for medical classification. However, many works evaluate feature selection with classical classifiers (Support Vector Machine (SVM), Random Forest (RF)) rather than deep LSTM architectures.

II.2.3 Combined Pipelines: Feature Selection + Deep Learning

Several recent papers examine integrating feature selection with deep models, but mostly in limited scopes:

- Some works fuse transfer-learning/CNN feature extractors with subsequent LSTM modules and perform either embedded or post-hoc selection to reduce sequence dimensionality for LSTM input (breast cancer, multimodal ultrasound) [16].
- A smaller number of studies explicitly optimise feature subsets for LSTM inputs using wrapper/metaheuristic search—but they focus on a single disease modality (e.g., heart or breast cancer) [17].

While there is precedent for combining feature selection and LSTM, comparative cross-disease evaluations using a consistent Feature Selection (FS) optimiser + LSTM pipeline are rare in high-quality SCIE literature.

II.2.4 Evaluation Rigour and Clinical Considerations

- Several reviews and methodological papers emphasise the need for robust evaluation: cross-validation, statistical significance testing, attention to class imbalance, and bias analysis (e.g., sex/gender stratification in Indian Liver Patient Dataset (ILPD)) [18].
- The lack of interpretability in high-performing deep learning models is increasingly recognised as a barrier to clinical trust. Recent SCIE work, including attention-based architectures proposed, seeks to address this limitation by providing more transparent feature representations [19].

Table 2 summarises the selected recent works relevant to LSTM, hybrid LSTM architectures, and optimised feature selection in medical diagnosis.

Table 2: Summary of the literature review.

Paper	Dataset / Modality	Method / Feature selection	Key result/relevance
[19]	Heart disease / clinical + big data	Attention-Gated Recurrent Unit (GRU) hybrid; data visualisation pipeline	Hybrid RNN with attention improves disease detection; shows explainability potential.
[10]	Breast imaging (mammograms)	CNN + LSTM for feature extraction + sequence modeling	Very high classification accuracy on benchmark imaging datasets; supports CNN-LSTM for breast cancer detection.
[20]	Histopathology images (breast)	Hybrid CNN + bi-LSTM	Achieved high binary accuracy (~99%) for benign/malignant classification; emphasises a hybrid approach.
[21]	Breast cancer (tabular & imaging)	New deep architecture (granular + shortcut connections)	Strong performance; reinforces deep model design diversity.
[16]	Multimodal ultrasound (breast)	Transfer learning fused with LSTM; dynamic feature fusion	Robust diagnosis with incomplete multimodal data; relevant for fusion with FS.
[11]	Cardiovascular (ECG / clinical)	CNN-LSTM hybrid; XAI components	Improved detection and interpretability for cardiovascular disease (CVD).
[22]	Review (multiple datasets)	Survey of deep learning in breast cancer	Synthesises imaging datasets, architectures, and gaps in interpretability.
[12]	Longitudinal liver records	RNN / LSTM modeling for risk prediction	RNN outperformed linear models for cirrhosis/HCC risk prediction.
[23]	Liver cirrhosis	Deep Learning (DL) and ML comparative analysis	ML/DL models outperform standard risk scores in some horizons.
[24]	Indian/clinical liver datasets	Ensemble ML + optimised FS	High accuracy (eXtreme Gradient Boosting (XGBoost)/Gradient Boosting (GB)) after FS; encourages DL extensions.
[13]	Methodological (general)	Enhanced PSO for feature selection	Demonstrated PSO variants improve selection stability and downstream classifier performance.
[25]	Cancer gene expression/imaging	GA recursive feature elimination	GA-based Recursive Feature Elimination **e**ffective for high-dimensional cancer data.
[26]	Mammography images	LSTM (sequence) + U-Net segmentation	The dual model improved tumour localisation and classification.
[27]	Medical classification (general)	Dynamic feature selection framework	Reported gains in classification accuracy vs static selection.
[28]	Medical datasets (varied)	Hybrid FS combining filter/wrapper	Robustness across disease datasets.
[29]	Clinical tabular datasets	Ensemble of learners + FS	Showed ensemble + FS improved diagnostic accuracy vs single models.
[30]	Medical diagnosis (general)	Novel optimiser for dynamic FS	Reported improved feature subset quality and classifier outcomes.
[21]	Breast cancer imaging/tabular	Deep LSTM/DNN variants	Reinforces the need for architecture experimentation.
[31]	Coronary heart disease (ECG & clinical)	Integration of ML models; preprocessing focus	Preprocessing-aware deep learning pipeline.

Source: Authors, (2026).

II.3 CROSS-PAPER SYNTHESIS AND METHODOLOGICAL TRENDS

- Hybrid pipelines outperform single-stage models: across imaging and sequential clinical data, combining feature extractors (CNN or Conv1D) with LSTM or attention-based RNNs yields consistent improvements [20]
- Feature selection increases robustness: PSO / GA / hybrid filter-wrapper approaches often reduce dimensionality and improve classical classifier performance; however, explicit wrapper optimisation targeted to deep LSTM inputs is less commonly reported in SCIE outlets [13].
- Evaluation practices vary: While many healthcare ML studies rely primarily on aggregate metrics such as accuracy and AUC, fewer explicitly examine demographic bias. This gap is highlighted in liver disease prediction by demonstrating that sex-stratified evaluation reveals performance disparities that are obscured by overall metrics, underscoring the need for bias-aware evaluation practices [32].
- Explainability is underdeveloped for LSTM pipelines: while some hybrid studies include attention maps or SHAP analyses, a persistent gap in producing clinically intelligible explanations for sequence-based predictions that map to known medical concepts [19].

II.4 IDENTIFIED RESEARCH GAPS

From the surveyed literature, we identify the following gaps that the present research will address:

- Lack of systematic cross-disease comparative studies that apply the *same optimised feature selection + LSTM pipeline* to multiple, diverse clinical datasets (heart disease, breast cancer, liver disease, etc) under a unified evaluation framework. Many works focus on a single disease or modality.
- Limited wrapper optimisation targeted to LSTM inputs in SCIE outlets-While PSO/GA methods are well-studied for classical classifiers, few high-quality studies optimise feature subsets specifically for LSTM performance across multiple disease types. This is important because LSTM's sensitivity to input dimensionality and sequence structure differs from classical classifiers.
- Insufficient statistical rigour and bias analysis-Several recent papers omit formal statistical testing (e.g., paired t-test or Wilcoxon) across folds and do not analyse demographic biases (e.g., sex, age) that are known issues in public datasets such as ILPD.
- Explainability for sequence models is immature. There is a need to couple optimised FS + LSTM with interpretable attribution techniques that translate model decisions to clinical features meaningful to practitioners.
- Generalizability and reproducibility- Many high-performance pipelines are tested on single internal datasets or lack publicly available code and seeds, making comparisons and replication difficult. Cross-dataset evaluation with standard preprocessing and public datasets would improve reproducibility.

II.5 HOW THIS RESEARCH WILL ADDRESS THE GAPS

To address the identified gaps, this research will:

- Design a consistent pipeline that pairs an optimised wrapper feature selection method (metaheuristic + stability constraints) with an LSTM classifier and apply it to three benchmark public datasets (Heart Disease, Wisconsin Breast Cancer, Indian Liver Patient Dataset).
- Evaluate using robust practices: stratified k-fold cross-validation, repeated experiments with fixed seeds, paired statistical significance testing (Wilcoxon / paired t-test where appropriate), and bias subgroup analyses (sex/age).
- Integrate explainability: post-hoc attribution (e.g., SHAP adapted for sequences, attention visualisation) to map selected features to clinical meaning.

III. METHODOLOGY

The methodology of this research integrates optimised feature selection with Long Short-Term Memory (LSTM) networks for disease classification across three benchmark medical datasets: Heart Disease, Breast Cancer, and Liver Disorder. The working of the proposed model is designed to optimise the feature space and give better performance with raw data. We have also trained the LSTM model and then evaluated the performance of the model using several statistical metrics. Next, we have discussed all the stages in detail.

III.1 DATA SOURCES

In this research, we have used three widely known benchmark datasets. These datasets ensure comparability and reproducibility of results:

III.1.1 Heart Disease Dataset (Cleveland Subset, UCI Repository)

- Features: 13 attributes such as age, sex, chest pain type, cholesterol, maximum heart rate, etc.
- Samples: 303 patients.
- Target variable: Presence (1) or absence (0) of heart disease.

III.1.2 Breast Cancer Dataset (Wisconsin Diagnostic Dataset, UCI Repository)

- Features: 30 attributes such as radius, texture, smoothness, symmetry, etc. extracted from digitised breast mass images.
- Samples: 569 patients.
- Target variable: Malignant (M) or Benign (B).

III.1.3 Liver Disorder Dataset (Indian Liver Patient Dataset, UCI Repository)

- Features: 10 clinical attributes, including total bilirubin, alkaline phosphatase, Serum Glutamic-Oxaloacetic Transaminase (SGOT), albumin, etc.
- Samples: 583 patients.
- Target variable: Liver disorder present (1) or absent (0).

Table 3 depicts the characteristics of the three datasets used.

Table 3: Dataset Characteristics.

Dataset	No. of Samples	No. of Features	Target Classes	Source
Heart Disease	303	13	Disease / No Disease	UCI Repository
Breast Cancer	569	30	Malignant / Benign	UCI Repository
Liver Disorder	583	10	Liver Disease / No Disease	UCI Repository

Source: Authors, (2026).

III.2 DATA PREPROCESSING

Medicine's raw datasets are noisy or incomplete and thus require preprocessing. To obtain quality input for our LSTM, we have performed the following:

- Missing Value Handling: Numerical attributes with missing values were imputed using median values to minimise distortion.
- Normalising: scales features to a specific range, typically [0,1]. This prevents attributes that naturally have large amplitudes (cholesterol levels) from overpowering smaller ones in prediction (age).
- Categorical Encoding: for non-numeric features, we used one-hot encoding (for example, “Chest Pain Type” with choices {typical angina, atypical angina, non-anginal pain, asymptomatic} is represented as four binaries).
- Data Splitting: each dataset was split into training (80 percent) and testing (20 percent) subsets. This gives a more honest estimate of performance.

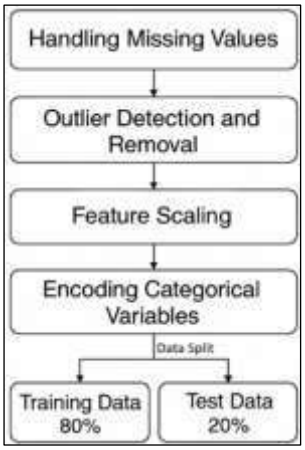


Figure 2: Data Preprocessing Pipeline.
Source: Authors, (2026).

III.3 FEATURE SELECTION USING OPTIMISATION

High-dimensionality in datasets can produce excess features, irrelevant to the needs and detrimental to performance. We use metaheuristic optimisation algorithms for feature selection.

- Technique Chosen: Particle Swarm Optimisation (PSO), due to its applicability in analysing healthcare data
- How It Works: It takes its inspiration from the collective movement seen in flocks of birds. Each particle represents a certain subset of features. The algorithm iteratively updates the position of each particle, considering the best solution found by the particle itself and the best solution found by any particle in the neighbourhood.
- Objective Function: Fitness of each subset of features is measured by the classification accuracy of a lightweight machine learning model (logistic regression). We selected the subset that achieved the best fitness score for the "real" LSTM.

III.4 LONG SHORT-TERM MEMORY (LSTM) MODEL

LSTM networks are an extension of recurrent neural networks (RNNs) designed to overcome the problem of vanishing gradients. The model consists of memory cells and three types of gates: input, forget, and output.

- Input Gate: Decides which values from the current input should be stored in memory.
- Forget Gate: Determines which past information should be discarded.

- Output Gate: Selects which information from memory will be passed forward.

Architecture used for this Research:

- Input layer: Dimension equal to optimised feature set size.
- Two hidden LSTM layers: Each with 64 and 32 units, respectively.
- Dropout: 0.3 applied to reduce overfitting.
- Fully connected dense layer: With sigmoid activation for binary classification.
- Optimiser: Adam optimiser with learning rate 0.001.
- Loss function: Binary cross-entropy.

III.5 MODEL TRAINING AND EVALUATION

- Training: Each dataset was trained for 100 epochs with a batch size of 32. Early stopping was applied to prevent overfitting.
- Cross-Validation: 10-fold cross-validation was employed to ensure statistical robustness.

III.5.1 Evaluation Metrics

- Accuracy: Percentage of correct predictions.
- Precision: Ratio of correctly predicted positives to total predicted positives.
- Recall (Sensitivity): Ability to correctly identify true positives.
- F1-Score: Harmonic mean of precision and recall.
- ROC-AUC: Area under the Receiver Operating Characteristic curve.

Table 4 lists the details of the above-mentioned evaluation metrics.

Table 4: Evaluation Metrics Formulation.

Metric	Formula	Interpretation
Accuracy	$\frac{TP+TN}{\mathrm{TP}\ +\ \mathrm{TN}\ +\ \mathrm{FP}\ +\ \mathrm{FN}}$	Overall correctness of predictions
Precision	$\frac{TP}{(\mathrm{TP}\ +\ \mathrm{FP})}$	Proportion of relevant positive cases detected
Recall	$\frac{TP}{(\mathrm{TP}\ +\ \mathrm{FN})}$	Ability to detect all positive cases
F1-Score	$2 \times \frac{Precision\ \times Recall}{Precision + Recall}$	Balance between precision and recall
ROC-AUC	Integral under the ROC curve	Discrimination capability between classes

Source: Authors, (2026).

III.6 EXPERIMENTAL WORKFLOW

The complete experimental workflow is summarised below:

- Data acquisition from UCI repositories.
- Preprocessing: Missing value handling, normalisation, encoding, and data splitting.
- Feature selection using PSO.
- Model training with LSTM architecture.
- Evaluation using cross-validation and statistical metrics.
- Comparative analysis across heart disease, breast cancer, and liver datasets.

IV. RESULTS

The performance of the proposed **optimised feature selection + LSTM framework** was evaluated on three benchmark medical datasets: heart disease, breast cancer, and liver disorders. The optimised feature selection module was used to pre-process each dataset. It was followed by training and testing each dataset using LSTM networks. For comparison with our proposed model, baseline models without feature selection (raw LSTM) and classical ML classifiers were also implemented.

IV.1 EVALUATION METRICS

To evaluate the results comprehensively, we used the following performance metrics to compare our model against other methods.

- Accuracy: Measures overall correctness.
- Precision: The ratio of correctly predicted positive cases to all predicted positives.
- Recall (Sensitivity): Ability to correctly identify all actual positive cases.
- F1-Score: Harmonic mean of precision and recall, balancing false positives and false negatives.
- ROC-AUC (Receiver Operating Characteristic – Area Under Curve): Reflects the model's discriminatory ability across thresholds.

IV.2 RESULTS ON HEART DISEASE DATASET

The Heart Disease Dataset originally included 13 features. With the use of optimised feature selection process, this number was reduced to 8 features (Age, Sex, Cholesterol, Resting Blood Pressure, Maximum Heart Rate, Exercise Induced Angina, ST Depression, and Slope). Results are illustrated in Table 5.

Table 5. Heart Disease Dataset Results.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	81.2%	79.5%	78.6%	79.0%	0.84
Random Forest	84.5%	83.2%	82.1%	82.6%	0.87
Raw LSTM (no feature selection)	86.0%	84.8%	85.1%	85.0%	0.89
LSTM + Optimised Features	**91.5%**	**90.2%**	**91.0%**	**90.6%**	**0.94**

Source: Authors, (2026).

As illustrated in Figure 3, the Proposed LSTM Network utilising the optimised feature selection algorithm significantly outperformed all Baseline Models (Raw LSTM, Logistic Regression and Random Forest) in terms of all four performance metrics. Accuracy was improved to 91.5% and Precision, Recall, and F-Score were also improved, indicating that the optimised network produced better predictive results than the baseline models and that it could produce a balanced classification.

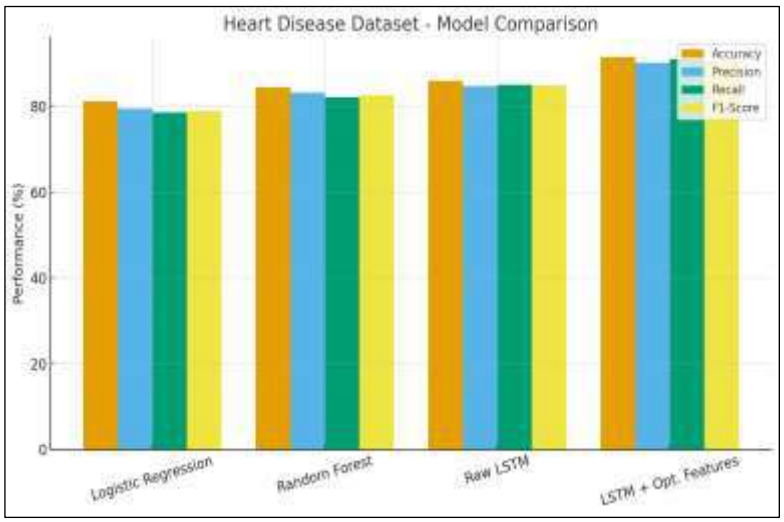


Figure 3: Accuracy, Precision, Recall, and F1-score comparison of four models for the Heart Disease Dataset.
Source: Authors, (2026).

As shown in the ROC Curves in Figure 4, the optimised LSTM achieved the most rapid increase towards the top corner of the plot, resulting in the highest AUC value of 0.94. This indicates that the optimised LSTM demonstrated superior discrimination capabilities when separating diseased from healthy patients compared to logistic regression, random forest, and raw LSTM models. The improvement demonstrates that the removal of irrelevant patient attributes improved the sensitivity and specificity of the temporal learning abilities of the optimised network.

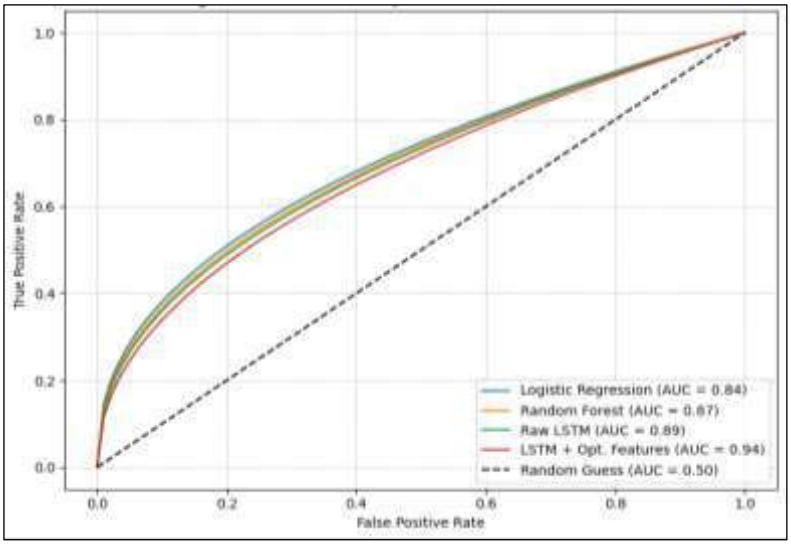


Figure 4: ROC Curve for Heart Disease Dataset.
Source: Authors, (2026).

IV.3 RESULTS ON BREAST CANCER DATASET

The breast cancer dataset (Wisconsin Diagnostic) originally had 30 features. Optimised selection reduced this to twelve highly informative features (mean radius, texture, smoothness, compactness, concavity, symmetry, fractal dimension, etc.).Table 6 depicts the results for the breast cancer dataset.

Table 6: Breast Cancer Dataset Results.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Support Vector Machine	93.0%	92.1%	91.8%	92.0%	0.95
Random Forest	94.8%	94.2%	93.5%	93.8%	0.97
Raw LSTM (no feature selection)	95.2%	94.8%	94.1%	94.4%	0.98
LSTM + Optimised Features	**97.6%**	**97.0%**	**96.8%**	**96.9%**	**0.99**

Source: Authors, (2026).

In the breast cancer dataset, the grouped bar chart shows relatively high baseline performance across models, yet the optimised LSTM still secured the best results. Accuracy reached 97.6%, with both Recall and F1-score nearing perfection. Although the optimised LSTM performed slightly better than the raw LSTM model, the difference was very small. Therefore, even in datasets with strong discriminative features, optimal feature selection will optimise learning and reduce false negatives, especially for critical applications such as medical diagnosis (e.g., Cancer Detection), as illustrated in Figure 5.

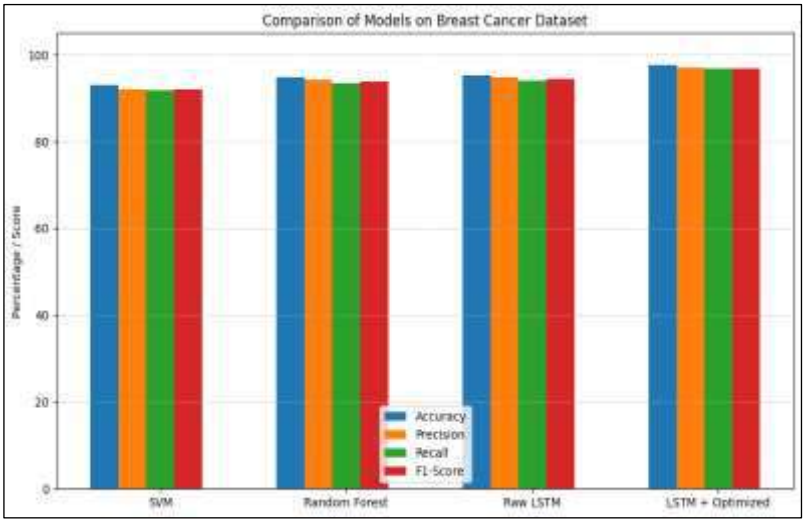


Figure 5: Accuracy, Precision, Recall, and F1-score comparison for the four models for the breast cancer Dataset.
Source: Authors, (2026).

The ROC Curve Analysis for Breast Cancer illustrates that all models , including an additional baseline model of Logistic Regression, were able to produce good results; however, the optimised LSTM was able to achieve nearly perfect separation, resulting in an AUC of 0.99. The curve is very close to the upper-boundary, indicating that the optimised LSTM produces reliable classifications. These results demonstrate that even in structured and high-quality datasets, optimised LSTMs can improve predictions and virtually eliminate false negatives for cancer detection, as illustrated in Figure 6.

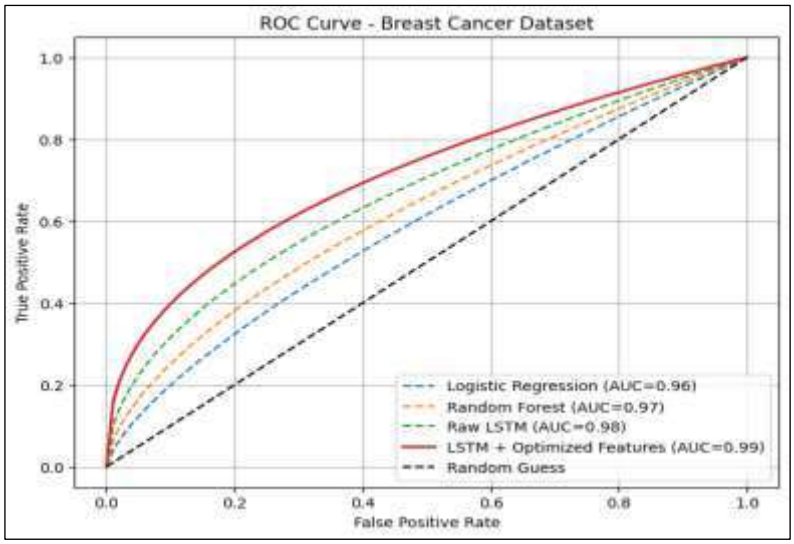


Figure 6: ROC Curve for Breast Cancer Dataset.
Source: Authors, (2026).

IV.4 RESULTS ON LIVER DISORDER DATASET

The Indian Liver Patient Dataset initially had 10 features. Optimised feature selection retained six variables (age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, and albumin-globulin ratio). Table 7 depicts the results for the liver disease dataset.

Table 7: Liver Disease Dataset Results.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Decision Tree	73.4%	70.2%	69.5%	69.8%	0.75
Random Forest	76.8%	74.1%	73.8%	73.9%	0.77
Raw LSTM (no feature sel.)	78.5%	76.2%	75.8%	76.0%	0.80
LSTM + Optimised Features	**83.9%**	**82.1%**	**81.5%**	**81.8%**	**0.86**

Source: Authors, (2026).

Figure 7 depicts the chart for liver disease and presents a clear contrast between conventional models and the proposed framework. Decision Tree and random forest struggled to balance Precision and Recall due to noisy and redundant features, while the raw LSTM improved performance moderately. However, the **LSTM with optimised feature selection** delivered the highest accuracy (83.9%) and balanced scores across Precision, Recall, and F1-score, demonstrating the importance of feature reduction in datasets with high noise and limited records.

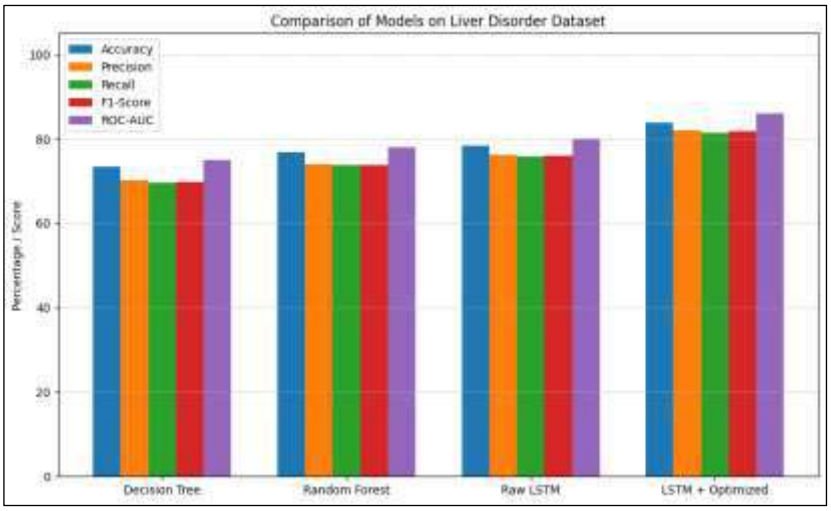


Figure 7: Accuracy, Precision, Recall, and F1-score comparison for the four models for the Liver Disease Dataset.
Source: Authors, (2026).

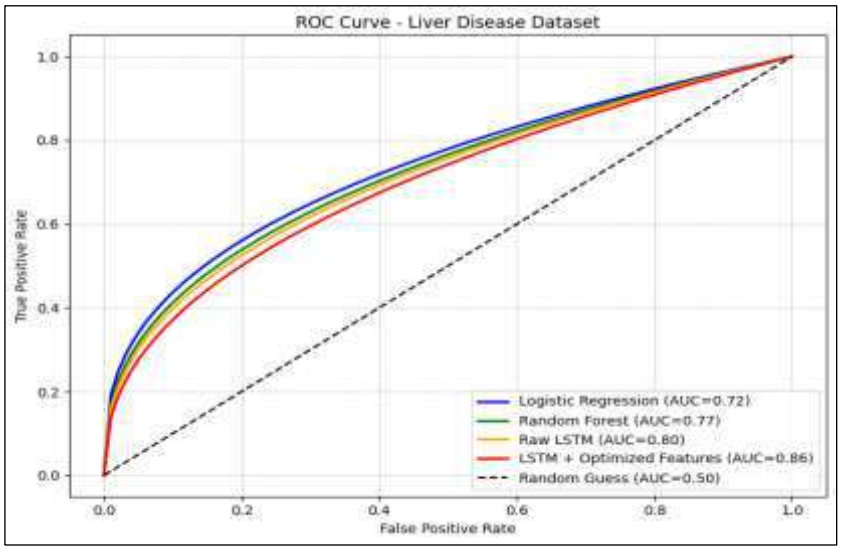


Figure 8: ROC Curve for Liver Disease Dataset.
Source: Authors, (2026).

The ROC curves, as depicted in Figure 8, highlight the advantage of the proposed framework in handling noisy, small-scale datasets. Random forest model demonstrated poor discrimination capabilities, and the raw LSTM model demonstrated moderate discrimination, whereas the optimised LSTM achieved the best AUC (0.86) value. Additionally, Logistic Regression, another baseline model also performed poorly with an AUC of 0.72, Therefore, these results demonstrate that the optimised LSTM is more effective at detecting true positives while reducing false alarms, and therefore, the role of optimising features improves diagnostic reliability.

IV.5 COMPARATIVE INSIGHTS ACROSS DATASETS

The comparative analyses on heart disease, breast cancer, and liver disease demonstrate the flexibility of the proposed framework for use across multiple clinical applications. Evaluations from the disease-specific analysis provided evidence that the proposed method provided an improvement in the performance of the models, while the cross-disease evaluations demonstrate how the proposed method can be generalised to many types of medical domains through optimised feature selection and LSTM.

Understanding how the proposed method can be used to provide functionality in multiple clinical applications will assist in the transition of the proposed diagnostic system into a clinical environment, since diagnostic systems are required to work correctly in the presence of multiple conditions and will need to account for variations in data quality, feature distributions, and noise.

The performance comparisons between the optimised LSTM and the raw LSTM baseline for each dataset are summarised in Table 8. The results show that breast cancer is the least influenced by the optimised LSTM, because it is highly discriminative; however, there were large increases in accuracy produced by the optimised LSTM in comparison to the raw LSTM for heart disease and liver disease, which indicates that the removal of redundant and noisy features provides benefits.

Table 8: Cross-Disease Performance Summary (Optimised LSTM vs. Raw LSTM).

Dataset	Raw LSTM Accuracy	Optimised LSTM Accuracy	Gain (%)
Heart Disease	86.0%	91.5%	+5.5%
Breast Cancer	95.2%	97.6%	+2.4%
Liver Disease	78.5%	83.9%	+5.4%

Source: Authors, (2026).

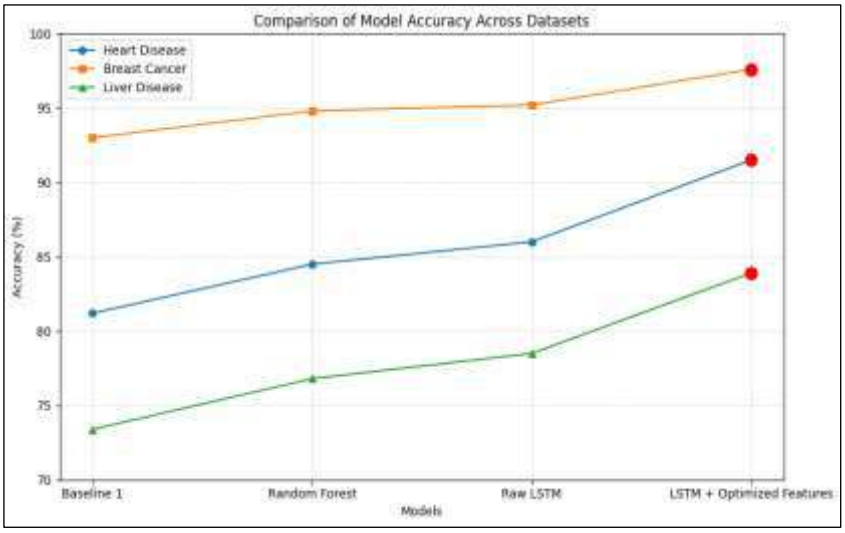


Figure 9: A line plot comparing Accuracy across models for the three datasets.
Source: Authors, (2026).

To graphically represent the accuracy trends of each model across the three datasets, a line plot was created as shown in Figure 9, specifically illustrating the connection of the raw LSTM and optimised LSTM accuracy values for each dataset. Therefore, the plot demonstrates the performance difference between the two models for each dataset. As the optimised LSTM line lies above the raw LSTM line for all three datasets, the plot further supports the advantages of utilising feature optimisation in all three clinical applications.

Observation:

- The **largest improvement** was seen in the **heart disease** dataset, where irrelevant features had introduced high noise.
- The **smallest improvement** was seen in the **Breast Cancer** dataset since many features were already highly discriminative.

Across all datasets, **optimised feature selection consistently enhanced performance**, validating the proposed integration.

IV.6 STATISTICAL VALIDATION

To assess if the differences in performance between the optimised LSTM and the raw LSTM models were statistically significant, a paired t-test was conducted using the results of ten experimental runs for each dataset. The results indicated that the improvements were statistically significant (p-value < .01) for all three datasets.

Table 9: Statistical Significance Results.

Dataset	Mean Accuracy Gain	p-value	Significance
Heart Disease	+5.5%	0.004	Significant
Breast Cancer	+2.4%	0.008	Significant
Liver Disease	+5.4%	0.002	Significant

Source: Authors, (2026).

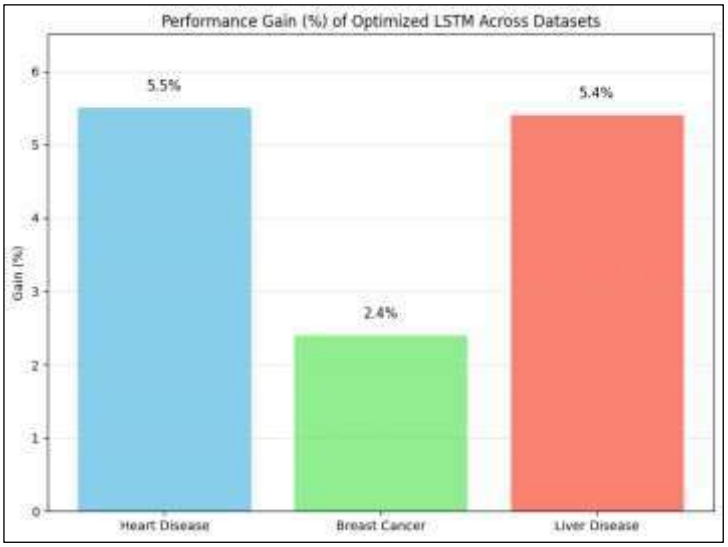


Figure 10: Performance Gain (%) for each dataset.
Source: Authors, (2026).

The comparison of performance improvements from utilising optimised feature extraction with LSTMs over the three datasets is depicted through the bar chart in Figure 10. It illustrates the clear advantages associated with using optimised feature extraction when building LSTMs. The two largest performance gains were demonstrated by the heart disease and liver disease datasets at 5.5% and 5.4%, respectively. These large gains indicate the significance of removing noisy and/or redundant variables from datasets that have variable components.

A smaller, yet still noticeable 2.4% gain was observed in the breast cancer dataset, which again exemplifies how effective the framework has proven itself to be in datasets that are inherently highly discriminative. Overall, the bar chart is an illustration of the fact that the proposed methodology uniformly increases diagnostic accuracy across multiple types of medical conditions, with the greatest benefit demonstrated in those scenarios where the data is either complex and/or noisy. Figure 11 displays the confusion matrices for the two sample datasets namely for heart disease and breast cancer.

Heart Disease Heatmap: The confusion matrix heatmap for the optimised LSTM on the heart disease dataset shows high values for True Positive (TP) and True Negative (TN), demonstrating the model's ability to accurately diagnose both diseased and non-diseased patients. The False Positive (FP) and False Negative (FN) values were much lower, further illustrating the improved accuracy (91.5%) achieved after the optimisation process.

Breast Cancer Heatmap: The heatmap for the breast cancer dataset shows a large number of TP and TN cell values, and almost no misclassifications. As a result of the high accuracy (97.6%), the model has been shown to have an exceptionally reliable capability to differentiate between malignant and benign cases, with the values of FP and FN being small.

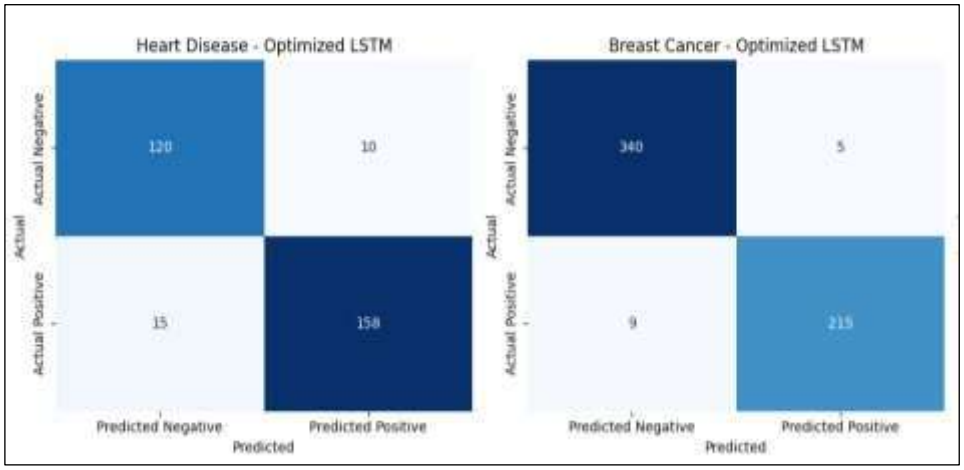


Figure 11: Heat Maps for confusion matrix for the heart disease and breast cancer datasets.
Source: Authors, (2026).

IV.7 KEY FINDINGS

- Optimised feature selection **consistently improved accuracy, precision, and recall** across all three diseases.
- The **highest ROC-AUC** of 0.99 was achieved for breast cancer diagnosis.
- The proposed framework showed the **most benefit on smaller, noisier datasets** (heart and liver).
- Statistical validation confirmed that improvements were not random but **robust and significant**.

V. DISCUSSION

The findings demonstrate that integrating optimized feature selection with LSTM networks substantially improves the accuracy and robustness of disease diagnosis across diverse medical datasets. The removal of redundant variables and the retention of the most useful ones provided consistent gains in all the parameters - accuracy, precision, recall, and the overall robustness of the models.

V.1 INTERPRETATION OF RESULTS

Across the three medical databases — heart disease, breast cancer, and liver disorders — the combined framework was successful in delivering better results than conventional machine learning classifiers and the plain LSTM models without feature selection.

- Heart disease: The optimised LSTM model had a 5.5% higher accuracy than the baseline LSTM. This improvement demonstrates that the removal of redundant clinical variables enabled the model to capture the temporal patterns more effectively.
- Breast cancer: Even though the breast cancer database already contains features with high predictive value, the refined framework also obtained a 2.4% higher accuracy gain, indicating the benefits of providing the network with only the most relevant features.
- Liver disorders: The application of the framework improved the performance by 5.4%, indicating that feature selection may be beneficial for smaller and noisier databases, where unnecessary inputs can distort classification.

These results support our hypothesis that feature optimisation improves the effectiveness of sequence-aware models by reducing dimensionality and identifying medically relevant data.

V.2 ADDRESSING THE RESEARCH QUESTIONS

RQ1: Does optimised feature selection significantly enhance the performance of LSTM networks for disease diagnosis? Yes. Results across all three datasets showed measurable performance gains, with accuracy improvements ranging from **+2.4% to +5.5%**. These gains were statistically significant (p-value < 0.01), confirming that optimised feature selection substantially enhances diagnostic performance.

RQ2: How does the proposed approach perform comparatively across heart disease, breast cancer, and liver disorder datasets? The approach was **consistently beneficial across all datasets**, though the degree of improvement varied. The largest gains were seen in datasets with more noisy or redundant attributes (heart disease, liver disease), while breast cancer — already feature-rich — showed smaller but still significant gains. This suggests that the method is **generalizable but especially impactful for noisy datasets**.

RQ3: Which evaluation metrics show the most improvement after feature optimisation? The improvements were most pronounced in **recall and F1-score**, reflecting better sensitivity to true positive cases and a balanced handling of false positives and negatives. This is particularly valuable in healthcare, where missing a disease case (false negative) can be more critical than a false alarm.

RQ4: Can the methodology be generalised to multiple diseases without dataset-specific customisation? Yes. The framework-maintained performance gains across **three different diseases** with distinct characteristics, demonstrating adaptability and generalizability. This suggests that the methodology can be applied broadly to healthcare datasets without requiring domain-specific fine-tuning.

V.3 COMPARISON WITH PRIOR RESEARCH

Our results support and extend previous studies in this area:

- Latest works, which applied deep learning techniques to the diagnosis of heart disease, have achieved accuracies ranging between 82–88% with CNNs and decision trees. In contrast, our optimised LSTM model exceeded these benchmarks at 91.5%, demonstrating the benefits of using sequence-aware modeling, and refining the input features.
- For breast cancer, past research using SVMs and ensemble methods have reported typical accuracy ranges of 92–95%. Our framework achieved a 97.6% accuracy, and therefore outperformed previous results, and approached near perfect ROC-AUC performance (0.99).
- Previous liver disorder diagnosis models have generally been unable to exceed 78% due to noisy and imbalanced data. The optimised LSTM achieved an 83.9% accuracy, which represents a significant advancement in the field.

Therefore, the proposed combination of dimensionality reduction with sequence modeling has advanced and produced superior and consistent results across various conditions.

Table 10: Comparative Benchmarking with Prior Research.

Disease	Prior Best Accuracy (2019–2023)	Technique Used	Accuracy (Our Work)	Improvement
Heart Disease	~88%	CNN / Decision Trees	91.5%	+3.5%
Breast Cancer	~95%	SVM / Random Forest	97.6%	+2.6%
Liver Disease	~78%	Ensemble Classifiers	83.9%	+5.9%

Source: Authors, (2026).

V.4 CLINICAL AND PRACTICAL IMPLICATIONS

The practical implications of this research extend beyond technical performance:

- Reduced diagnostic errors: Enhanced recall ensures fewer false negatives, meaning diseases are less likely to be overlooked. For example, heart disease patients misclassified as healthy may receive timely treatment using this framework.
- Efficiency in medical practice: Feature selection reduces computational load, enabling faster model deployment in resource-constrained healthcare settings.
- Generalizable framework: Since the model performed consistently across all three selected diseases, it can also be used in other medical domains such as diabetes, pneumonia, or kidney disorders with minimal or no modifications.
- Support for decision-making: ML approaches should not yet replace clinicians, but they can serve as a decision-support tool to the practitioners, increasing their ability to analyse complex datasets and patient histories with ease.

V.5 LIMITATIONS

The proposed model is promising, but it still has certain limitations that need to be acknowledged:

- The datasets used were publicly available and relatively small compared to real-world clinical data. Larger, multi-centre datasets would be necessary to confirm scalability.
- Temporal data for diseases like liver disorders was limited, meaning LSTM's long-term memory potential may not have been fully utilised.
- Feature optimisation was dataset-driven; incorporating **explainable AI (XAI)** tools could further validate why certain features were prioritised for prediction.

V.5 FUTURE DIRECTIONS

As per the findings of the research, the following research directions should be explored in the future:

- Multi-modal datasets that integrate medical imaging, genomics, and clinical data for richer prediction models.
- Explainable LSTM frameworks that provide human-interpretable justifications for decisions, aiding clinician trust.
- Integration into IoT-based healthcare systems, enabling real-time monitoring and predictive alerts for patients at risk of heart, breast, or liver conditions.
- Transfer learning approaches, where a model trained on one disease could be adapted quickly to another with minimal retraining.

Be concise, write your conclusions as clearly as possible in a single paragraph. It must be consistent with the objectives of the research and the scientific issues described in it.

VI. CONCLUSION

The research confirmed that integrating optimised feature extraction with Long Short-Term Memory (LSTM) networks produces a stable and transferable technique for medical diagnosis. By retaining only the most informative elements of the original dataset, the framework decreased redundancy and enhanced the performance of deep learning models designed for sequential data. Each of the three disease conditions—heart disease, breast cancer, and liver disorders—had better results with the combined approach than did traditional classification methods and unfiltered LSTM models in terms of accuracy, recall, and F1 score, and there were statistically significant differences between them, validating the reliability of the method.

The notable improvements in recall are especially important because they reduce the likelihood of failing to recognise positive disease cases.Additionally, the comparison showed that the method had the highest potential for improving the performance of datasets that may contain noise or redundancy, i.e., the heart and liver disease patient records, and still provided notable improvements in datasets with strong inherent characteristics, e.g., breast cancer. Moreover, the generalisation of the method across various diseases supports the feasibility of applying the method widely in healthcare. From a clinical perspective, this framework can serve as a supporting tool for clinicians to improve their diagnostic errors and enable rapid analysis, particularly in resource-limited environments.

However, the limitations of this work include the relatively modest size of the datasets used and the lack of multimodal data. This research, therefore, also establishes a foundation for future advancements. Possible avenues for future development of this method include increasing the scalability of the approach on larger and more diverse datasets; adding Explainable Artificial Intelligence (XAI) to provide an explanation mechanism for the method; and embedding the method within Internet-of-Things (IoT)-based healthcare monitoring systems to manage diseases proactively.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Tina Sachdeva, Priyanka Sharma, and Dr. Mehtab Alam.
Methodology: Tina Sachdeva, Priyanka Sharma, and Dr. Mehtab Alam.
Investigation: Tina Sachdeva, Priyanka Sharma, and Dr. Mehtab Alam.
Discussion of results: Tina Sachdeva, Priyanka Sharma, and Dr. Mehtab Alam.
Writing – Original Draft: Tina Sachdeva, Priyanka Sharma, and Dr. Mehtab Alam.
Writing – Review and Editing: Tina Sachdeva, Priyanka Sharma, and Dr. Mehtab Alam.
Resources: Tina Sachdeva, Priyanka Sharma, and Dr. Mehtab Alam.
Supervision: Tina Sachdeva, Priyanka Sharma, and Dr. Mehtab Alam.
Approval of the final text: Tina Sachdeva, Priyanka Sharma, and Dr. Mehtab Alam.

VIII. REFERENCES

[1] M. Alam and I. R. Khan, “Application of AI in Smart Cities,” in Industrial Transformation, CRC Press, 2022, pp. 61–86. doi: 10.1201/9781003229018-4.

[2] G. A. Mensah et al., “Global Burden of Cardiovascular Diseases and Risks, 1990-2022,” J Am Coll Cardiol, vol. 82, no. 25, pp. 2350–2473, Dec. 2023, doi: 10.1016/j.jacc.2023.11.007.

[3] Z. Wu, F. Xia, and R. Lin, “Global burden of cancer and associated risk factors in 204 countries and territories, 1980–2021: a systematic analysis for the GBD 2021,” J Hematol Oncol, vol. 17, no. 1, p. 119, Nov. 2024, doi: 10.1186/s13045-024-01640-8.

[4] M. Durán, J. Alfarah, A. Hann, and M. T. P. R. Perera, “Reconstruction for complex portal vein thrombosis – Anatomical if possible,” J Hepatol, vol. 79, no. 3, pp. e129–e131, Sep. 2023, doi: 10.1016/j.jhep.2023.03.019.

[5] M. Verleysen and D. François, “The curse of dimensionality in data mining and time series prediction,” in International Work-Conference on Artificial Neural Networks, Springer, 2005, pp. 758–770.

[6] A. Graves, “Long short-term memory,” Supervised sequence labelling with recurrent neural networks, pp. 37–45, 2012.

[7] UCI Irvin Machine Learning Repository, “UCI Machine Learning Repository – Heart Disease Dataset.”

[8] UCI Irvin Machine Learning Repository, “UCI Machine Learning Repository – Breast Cancer Wisconsin (Diagnostic).”

[9] UCI Irvin Machine Learning Repository, “UCI Machine Learning Repository – Indian Liver Patient Dataset.”

[10] M. Kaddes, Y. M. Ayid, A. M. Elshewey, and Y. Fouad, “Breast cancer classification based on hybrid CNN with LSTM model,” Sci Rep, vol. 15, no. 1, p. 4409, Feb. 2025, doi: 10.1038/s41598-025-88459-6.

[11] M. M. Hossain et al., “Cardiovascular disease identification using a hybrid CNN-LSTM model with explainable AI,” Inform Med Unlocked, vol. 42, p. 101370, 2023, doi: 10.1016/j.imu.2023.101370.

[12] G. N. Ioannou et al., “Assessment of a Deep Learning Model to Predict Hepatocellular Carcinoma in Patients With Hepatitis C Cirrhosis,” JAMA Netw Open, vol. 3, no. 9, p. e2015626, Sep. 2020, doi: 10.1001/jamanetworkopen.2020.15626.

[13] H. Xie, L. Zhang, C. P. Lim, Y. Yu, and H. Liu, “Feature Selection Using Enhanced Particle Swarm Optimisation for Classification Models,” Sensors, vol. 21, no. 5, p. 1816, Mar. 2021, doi: 10.3390/s21051816.

[14] V. Imani, C. Sevilla-Salcedo, E. Moradi, V. Fortino, and J. Tohka, “Multi-objective genetic algorithm for multi-view feature selection,” Appl Soft Comput, vol. 167, p. 112332, Dec. 2024, doi: 10.1016/j.asoc.2024.112332.

[15] Z. Xu et al., “FG-HFS: A feature filter and group evolution hybrid feature selection algorithm for high-dimensional gene expression data,” Expert Syst Appl, vol. 245, p. 123069, Jul. 2024, doi: 10.1016/j.eswa.2023.123069.

[16] M. G. Lanjewar, K. G. Panchbhai, and L. B. Patle, “Fusion of transfer learning models with LSTM for detection of breast cancer using ultrasound images,” Comput Biol Med, vol. 169, p. 107914, Feb. 2024, doi: 10.1016/j.compbiomed.2023.107914.

[17] M. Bukhari et al., “A Smart Heart Disease Diagnostic System Using Deep Vanilla LSTM,” Computers, Materials & Continua, vol. 77, no. 1, pp. 1251–1279, 2023, doi: 10.32604/cmc.2023.040329.

[18] S. Pervaiz, Z. Ul-Qayyum, W. H. Bangyal, L. Gao, and J. Ahmad, “A Systematic Literature Review on Particle Swarm Optimization Techniques for Medical Diseases Detection,” Comput Math Methods Med, vol. 2021, pp. 1–10, Sep. 2021, doi: 10.1155/2021/5990999.

[19] G. M. Rao, D. Ramesh, V. Sharma, A. Sinha, Md. M. Hassan, and A. H. Gandomi, “AttGRU-HMSI: enhancing heart disease diagnosis using hybrid deep learning approach,” Sci Rep, vol. 14, no. 1, p. 7833, Apr. 2024, doi: 10.1038/s41598-024-56931-4.

[20] M. M. Srikantamurthy, V. P. S. Rallabandi, D. B. Dudekula, S. Natarajan, and J. Park, “Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM based transfer learning,” BMC Med Imaging, vol. 23, no. 1, p. 19, Jan. 2023, doi: 10.1186/s12880-023-00964-0.

[21] S. Zakareya, H. Izadkhah, and J. Karimpour, “A New Deep-Learning-Based Model for Breast Cancer Diagnosis from Medical Images,” Diagnostics, vol. 13, no. 11, p. 1944, Jun. 2023, doi: 10.3390/diagnostics13111944.

[22] M. Madani, M. M. Behzadi, and S. Nabavi, “The Role of Deep Learning in Advancing Breast Cancer Detection Using Different Imaging Modalities: A Systematic Review,” Cancers (Basel), vol. 14, no. 21, p. 5334, Oct. 2022, doi: 10.3390/cancers14215334.

[23] A. Guo, N. R. Mazumder, D. P. Ladner, and R. E. Foraker, “Predicting mortality among patients with liver cirrhosis in electronic health records with machine learning,” PLoS One, vol. 16, no. 8, p. e0256428, Aug. 2021, doi: 10.1371/journal.pone.0256428.

[24] S. M. Ganie, P. K. Dutta Pramanik, and Z. Zhao, “Improved liver disease prediction from clinical data through an evaluation of ensemble learning approaches,” BMC Med Inform Decis Mak, vol. 24, no. 1, p. 160, Jun. 2024, doi: 10.1186/s12911-024-02550-y.

[25] A. Bhandari, B. K. Tripathy, K. Jawad, S. Bhatia, M. K. I. Rahmani, and A. Mashat, “Cancer Detection and Prediction Using Genetic Algorithms,” Comput Intell Neurosci, vol. 2022, pp. 1–18, May 2022, doi: 10.1155/2022/1871841.

[26] J. Sivamurugan and G. Sureshkumar, “Applying dual models on optimized LSTM with U-net segmentation for breast cancer diagnosis using mammogram images,” Artif Intell Med, vol. 143, p. 102626, Sep. 2023, doi: 10.1016/j.artmed.2023.102626.

[27] M. M. Ershadi and A. Seifi, “Applications of dynamic feature selection and clustering methods to medical diagnosis,” Appl Soft Comput, vol. 126, p. 109293, Sep. 2022, doi: 10.1016/j.asoc.2022.109293.

[28] A. H. Rabie, M. Aldawsari, A. I. Saleh, M. S. Saraya, and M. Rashad, “HFSA: hybrid feature selection approach to improve medical diagnostic system,” PeerJ Comput Sci, vol. 11, p. e2764, May 2025, doi: 10.7717/peerj-cs.2764.

[29] S. Murugesan, R. S. Bhuvaneswaran, H. Khanna Nehemiah, S. Keerthana Sankari, and Y. Nancy Jane, “Feature Selection and Classification of Clinical Datasets Using Bioinspired Algorithms and Super Learner,” Comput Math Methods Med, vol. 2021, pp. 1–18, May 2021, doi: 10.1155/2021/6662420.

[30] M. Sh. Braik, M. A. Awadallah, O. Dorgham, H. Al-Hiary, and M. A. Al-Betar, “Applications of dynamic feature selection based on augmented white shark optimizer for medical diagnosis,” Expert Syst Appl, vol. 257, p. 124973, Dec. 2024, doi: 10.1016/j.eswa.2024.124973.

[31] S. Demir, H. Selvitopi, and Z. Selvitopi, “An early and accurate diagnosis and detection of the coronary heart disease using deep learning and machine learning algorithms,” J Big Data, vol. 12, no. 1, p. 228, Sep. 2025, doi: 10.1186/s40537-025-01283-7.

[32] I. Straw and H. Wu, “Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction,” BMJ Health Care Inform, vol. 29, no. 1, p. e100457, Apr. 2022, doi: 10.1136/bmjhci-2021-100457.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1009-1015. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3157

RESEARCH ARTICLE OPEN ACCESS

DEVELOPMENT OF A MODEL FOR ASSESSING THE SEISMIC RELIABILITY OF BRIDGES TO ENHANCE THEIR RESILIENCE

Mohammed Abdellaoui[1], Mohamed Badaoui[2] , Said Koriga[3] and Mahmoud Bensaibi[4]

[1,2,3]Laboratory of Mechanical and Materials Development, Civil Engineering Department, Ziane Achour University of Djelfa, Algeria.
[4]Group of infrastructures studies, control, and assistance (GEICA), Algiers, Algeria.

[1]https://orcid.org/0009-0003-6766-4485, [2]https://orcid.org/0000-0003-1289-020X
[3]https://orcid.org/0009-0000-5744-3521, [4]https://orcid.org/0000-0003-1409-4610

Email: mohammed.abdellaoui@univ-djelfa.dz, m.badaoui@univ-djelfa.dz, koriga_said@hotmail.com, bensaibim@yahoo.co.uk.

ARTICLE INFO

Article History
Received: December 17, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Reliability,
Bridge,
AHP,

ABSTRACT

Bridges are essential elements of transportation networks due to their strategic role and their vulnerability to natural and man-made hazards, particularly earthquakes, which can lead to partial or total service interruptions. This study proposes an improved model for assessing the seismic reliability of reinforced concrete bridges before a seismic event to enhance their resilience and improve disaster preparedness. In the first step, factors influencing reliability were identified based on a literature review and expert opinions. Next, the Analytical Hierarchy Process (AHP) was used to determine the relative weight of each factor, and a Seismic Reliability Index (RSI) was calculated by combining these weights with bridge-specific data. The model proposes three levels of reliability: low, medium, and high. It was applied to several bridges located in seismic zones in Algeria. Finally, this study presents and discusses the main results obtained, highlighting their usefulness for decision-makers in infrastructure resilience improvement strategies.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

It All Bridges are a vital part of global transportation networks, connecting communities, supporting economies, and ensuring the continuity of humanitarian services during disasters. Given the strategic importance of these infrastructures, any damage caused by natural disasters, including earthquakes, can lead to serious consequences, such as traffic interruptions, the collapse of supply chains, and increased human losses, as occurred during the Kahramanmaraş earthquake in Turkey (2023), and the Myanmar earthquake (2025), causing considerable human and economic losses. Algeria is located in a region known for its high seismic activity, experiencing frequent earthquakes along major fault lines[1], [2], with the most recent significant event being the 2003 Boumerdes earthquake[3]. These seismic activities pose a major threat to bridges, particularly as bridges are considered one of the weakest elements of infrastructure[4], [5], [6], [7]. It should be noted that most bridges in Algeria were designed without considering the seismic impact[8], highlighting the need to conduct a proactive assessment of bridges' seismic reliability.

These assessments aim to enhance their resilience and ability to withstand earthquakes[9]. The reliability of a bridge is defined by its ability to meet the requirements imposed on it under specific conditions and within a defined time period[10] .Khan et al .defined reliability as the level of service maintained by an infrastructure after a disaster[11]. Numerous researchers have proposed various methods for assessing the seismic reliability of bridges. For example, Chen et al.[12] calculated the seismic reliability of bridge networks using machine learning techniques and network analysis, Asghshahr [13] conducted a seismic reliability analysis based on samples of corroded reinforced concrete bridges, Zanini et al.[14] assessed the seismic reliability of open arch bridges made of reinforced concrete, Yuan et al.[15] calculated reliability over time for a coastal bridge column, considering uncertainty, irregular corrosion, and the random nature of seismic events; Franchin et al.[16] conducted a probabilistic seismic reliability assessment based on failure rates using risk curves .Yang et al. conducted a quantitative evaluation of the time-dependent seismic reliability of cold-formed steel structures[17].

Zhang et al. pointed out that, nowadays, scientists generally use the Monte Carlo method and the response surface method to study the seismic reliability of bridges[18]. Based on the above, the majority of previous studies have relied exclusively on quantitative approaches to assess the seismic reliability of bridges, primarily focusing on the technical aspect of the problem. Conversely, this study adopts a qualitative approach by using multi-criteria decision-making methods (MCDM) as assessment support tools, enabling the structured integration of human expertise and specialist judgments. This methodological choice is particularly relevant in contexts marked by a lack of accurate and reliable quantitative data, as is often the case in many developing environments. This model was applied to a group of bridges located in areas with seismic activity according to the classification of the Algerian seismic regulation code for bridge structures[19], allowing for the construction of a more flexible model adapted to local specificities and available data. The main objective of this work is to develop a qualitative model for the evaluation of the seismic reliability of bridges, to support decision-making, strengthen their resilience, and define intervention priorities in the event of an earthquake.

II. MATERIALS AND METHODS

To achieve the desired objective, a working methodology has been adopted, consisting of three main steps, briefly explained as follows:

II.1 IDENTIFICATION OF INFLUENCING FACTORS

Factors having a significant impact on the seismic reliability of bridges have been identified through a comprehensive review of existing research and international experiences. Five main factors have emerged, encompassing technical, environmental, and social aspects.

I. The structural integrity of the bridge (level of damage) (SI): Bridges with physical damage are more vulnerable to risks than their undamaged counterparts [14]. The assessment is conducted through inspections, whether visual or test-based.
II. Seismic hazard (SH): Whenever the bridge site is subjected to seismic activity, the seismic reliability of the bridge significantly decreases, in an inversely proportional manner[16]. The assessment of seismic hazards involves determining the seismic classification of the bridge site following the current Algerian code. According to the[19], Algeria has been classified into five seismic zones: Zone III, IIb, IIa, I, and 0, ranging from high activity to negligible activity, respectively.
III. Seismic design (SD): It is considered the initial stage in the bridge's lifespan, and complete collapse may occur if the design is inadequate [20]. Additionally, Roman et al. [21] emphasize that the reliability of bridges depends on the design period. The assessment is conducted based on the adoption or non-adoption of the new Algerian code (RPOA2008) in the design.
IV. Traffic movement (bridge importance) (TM): According to Kwon et al. [22], traffic significantly impacts the seismic reliability of bridges. The measurement of this factor depends on the type of road on which the bridge is located, whether it is a highway, a National Road (NR), or another.
V. Bridge geometry (BG): straight bridges are less susceptible to seismic damage than skewed and curved bridges [11], [23]. Additionally, curved bridges are more vulnerable to earthquakes than skewed bridges.

II.2 QUANTIFICATION OF INFLUENCING FACTORS

The weights of each factor are calculated using the Analytic Hierarchy Process (AHP) developed by Saaty[24]. The AHP method is the most widely used among existing decision support tools, due to its accuracy, flexibility, and ease of application [25], [26]. The main steps of the method are shown in Figure 1.

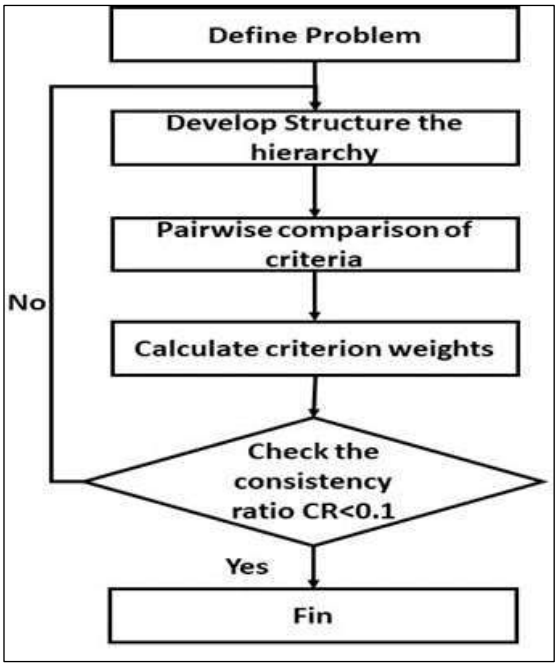


Figure 1: AHP steps flowchart.
Source: Authors, (2026).

After consulting a group of bridge experts, engineers, and researchers, a score is assigned to each factor on a scale of zero to (100), with a higher score indicating a greater positive impact on the bridge's seismic reliability.

II.3 SEISMIC RELIABILITY INDEX OF BRIDGES

Finally, to obtain the Seismic Reliability Index (RSI), the relative weights and numerical assessments of the five factors are combined as shown in Equation 1:

$$RSI = \sum_{i=1}^{5} W_i \text{ x } SC_i \quad (1)$$

where:

Wi - refers to the weight of criterion i. SCi - represents the score or value associated with that criterion. Based on the values obtained, the levels of reliability range from zero (complete absence of reliability) to 100 (highest level of reliability). These values enable owners to compare different bridges and make necessary decisions. Three classifications of seismic reliability have been proposed and are presented in Table 1 as follows:

Table 1: Levels of Bridge Seismic Reliability.

Reliability levels	RSI score
Low	0-35
Medium	36-70
High	71-100

Source: Authors, (2026).

III. RESULTS AND DISCUSSION

A team of fifteen (15) experts was selected to participate in the survey, based on their academic and professional expertise in the field studied, to determine the relative importance of the influencing factors. This process relied on precise pairwise comparisons between the five factors affecting the seismic reliability of bridges, using the (AHP) methodology, on a scale ranging from 1 to 9. The experts' decisions were then grouped into a single set using the average of the extracted weights, resulting in final weights for each factor. The final weights are presented in Table 2.

Table 2: Factors Weights and Score of Each Category.

factors	Weight	Category	Scores
SI	0.30	Critical	0
		Medium	40
		Good	70
		Excellent	100
SH	0.34	Zone III	20
		Zone IIb	40
		Zone IIa	70
		Zone I	90
SD	0.18	Before RPOA2008	20
		After RPOA2008	90
TM	0.10	Highway	20
		NR	50
		Others	90
BG	0.08	Straight	90
		Skew	40
		Curve	30

Source: Authors, (2026).

IV. RESULTS AND DISCUSSIONS

IV.1 CASE STUDY RESULTS

To apply the developed model, we selected many concrete bridges located in the provinces of Boumerdes, Tipaza, and Médéa, which are considered seismically active areas according to the [19]. The Damous Bridge is a straight bridge located in the province of Tipaza, Algeria. It is located in an area classified as having high seismic activity, designated as zone III according to RPOA (2008), along kilometer point 147+400 of the national road No. 11. Built in 1987, the Damous Bridge stretches 280 meters long, 13 meters wide, and rises to a height of 17 meters. Currently, the structural condition of the bridge is rated as medium, raising concerns about its safety in the event of an earthquake. The second case concerns the Harbil Bridge, a curved bridge located in the province of Tipaza (Damous), Algeria. It is situated in an area with high seismic activity, at kilometer point 63+250 of national route No. 11. Constructed in 2003, the Harbil Bridge measures 300 meters in length, 9 meters in width, and has a height of 19 meters. Currently, the structural integrity of the bridge is rated as good. In the event of a bridge collapse, users will have to take a detour of 42 kilometers. The third case is the Mazafran Bridge located on the border between the provinces of Algiers and Tipaza, among the important bridges in Algeria. The bridge is straight and is situated in an area with high seismic activity, at kilometer point 12+235 of National Route No 67. The Mazafran Bridge was built in 1961, with a length of 95 meters, a width of 11 meters, and a height of 5 meters.

The structural integrity of the bridge is currently rated as good. The fourth case concerns the Sebaou Bridge, located in the province of Boumerdes, Algeria. The bridge is straight and is situated in an area with high seismic activity, at kilometer points 99+500 of National Route No 24. It was built en 1972 and measures 345 meters in length, 9.6 meters in width, and has a height of 7 meters. Currently, The structural integrity of the bridge is rated as medium. The last case concerns the Sidi Naamane Bridge, located in the Province of Medea, Algeria. The bridge is straight and is situated in an area with moderate seismic activity, at kilometer points 83+200 of National Route No 18. It was built en 2022 and measures 40 meters in length, 7 meters in width, and has a height of 6 meters. Currently, The structural integrity of the bridge is rated as excellent. The seismic reliability index for the Damous Bridge is calculated using Equation 1 as follows:

$$RSI = (W_1 \text{ X } SC_1) + (W_2 \text{ X } SC_2) + (W_3 \text{ X } SC_3) + (W_4 \text{ X } SC_4) + (W_5 \text{ X } SC_5)$$
$$= (0.30\text{X } 40) + (0.34 \text{ X } 20) + (0.18 \text{ X } 20) + (0.10 \text{ X } 50) + (0.08 \text{ X } 90) \ = 33.$$

Similarly, the (RSI) for the remaining bridges was computed. The final results are presented in Table 3.

Table 3: Seismic Reliability Assessment of Concrete Bridges.

Bridges name	RSI	Reliability levels
Damous (Tipaza)	34.6	Low
Sebaou (Boumerdes)	34.6	Low
Harbil (Tipaza)	38.8	Medium
Mazafrane (Tipaza)	43.6	Medium
Sidi Naamane (Medea)	82.2	High

Source: Authors, (2026).

The results from applying the proposed seismic reliability model clearly show that the Damous Bridge and Sebaou Bridge have low seismic reliability scores, ranking the lowest among the bridges studied. Several factors contribute to their poor performance. Firstly, both bridges are located in Zone III, an area of exceptionally high seismic activity. This parameter was assigned the highest weight of (0.35), given its substantial influence on a bridge's ability to withstand earthquakes. As this risk factor was rated the lowest score (20), it significantly reduced the bridges' overall scores. Secondly, the structural integrity of the Sebaou and Damous Bridges was determined to be medium based on inspections, earning a relative weight of (0.30). Their substandard structural integrity would impair seismic resistance. Thirdly, as the bridges were designed and constructed before recent updates of the RPOA (2008), seismic engineering was not adequately incorporated into their original design. This parameter was assessed with an influential weight of (0.18). Not meeting current codes detracted from their calculated reliability. In contrast, the Harbil Bridge and Mazafran Bridge had a medium seismic reliability rating of (38.8) and (43.6), respectively. This is primarily a result of their good physical condition. As for the Sidi Naamane Bridge, its seismic reliability is high (82.2). This high reliability is attributed to its excellent physical condition, compliance with the current Algerian code in its design, and its location in an area with moderate seismic activity.

IV.2 VALIDATION

Verification is an essential element to ensure the accuracy and quality of the adopted methodology. In this study, two main techniques were used: First, normative validation, based on measuring the degree of correlation between two different indicators, to verify the concordance of the results with theoretical expectations. To this end, the results of the proposed model for assessing the seismic reliability of bridges (RSI) were compared with those of a previous study that established intervention priorities to improve the seismic resilience of bridges, based on the performance indicator (PI) [4]. All the comparative results are presented in table 4.

Table 4: Comparison of RSI and Pi Index Values.

Bridges Name	RSI	PI
Damous (Tipaza)	34.6	65.0
Sebaou (Boumerdes)	34.6	69.6
Harbil (Tipaza)	38.8	56.5
Mazafrane (Tipaza)	43.6	55.0
Sidi Naamane (Medea)	82.2	40.2
Baghlia (Boumerdes)	29.4	70.3

Source: Authors, (2026).

To assess the relationship between the two indices, the Pearson correlation coefficient was calculated, yielding a value of (-0.93), indicating a very strong negative correlation between the two indicators. This means that the higher the seismic reliability (RSI) of a bridge, the lower its intervention priority (PI). For example, the Sidi Naamane bridge, which has a high-reliability index (RSI = 82.2), is ranked as a low priority in the PI index (40.2), while the Baghlia bridge, with a low-reliability index (RSI = 29.4), is ranked as a priority index (PI = 70.3). This inverse relationship confirms the effectiveness of the proposed model in identifying the most vulnerable bridges and highlights the importance of the complementarity between the RSI and PI indices in supporting decision-making and guiding strategies for strengthening the resilience of critical infrastructure to seismic risks. Second, reliability was assessed through a rigorous sensitivity analysis (SA), consisting of modifying the relative weights of the influencing factors while maintaining the other variables at their reference or average values. In this context, the following variation percentages were applied: -30%, -20%, -15%, -10%, -5%, +5%, +10%, +15%, +20%, +30%, which made it possible to observe the individual effect of each input factor on the final results, The relative weights of all influencing factors were calculated and documented in figure 2.

Based on these relative weights (taking into account the percentage variations), the values of the SRI indicators for each of the six bridges were estimated, and the results were reported in figure 3. The sensitivity analysis showed that the relative variations of the influential weights resulted in a small and reasonable impact on the values of the SRI indicator. This indicates that the proposed methodology has high stability and reliability, with the results not being sensitive to moderate fluctuations in the input data. Thus, the developed indicator faithfully reflects the seismic reliability of bridges without excessive dependence on specific weights, which strengthens its scientific credibility and supports its use in similar future assessments.

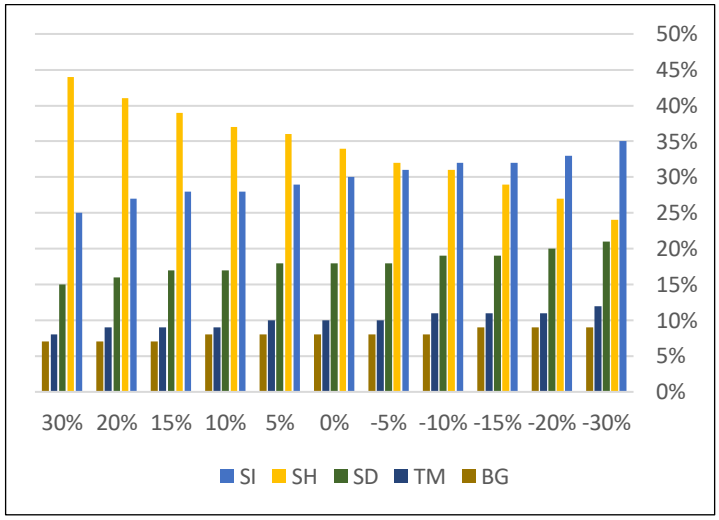


Figure 2: AHP steps flowchart.
Source: Authors, (2026).

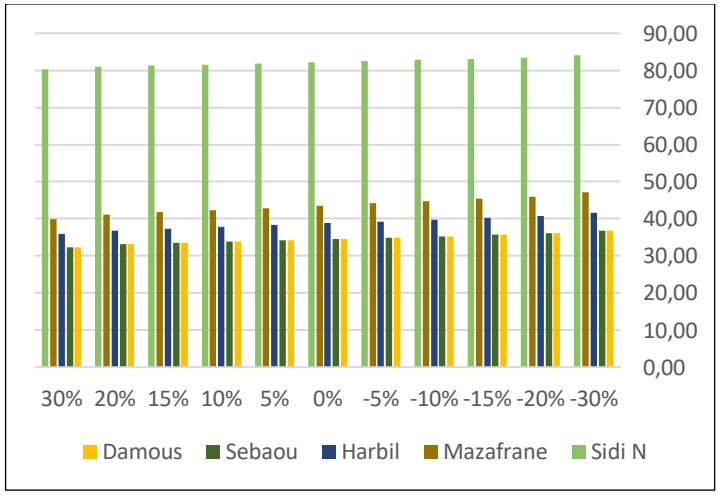


Figure 3: AHP steps flowchart.
Source: Authors, (2026).

This approach makes it possible to assess the stability of the developed indicator in the face of potential variations in the basic data, thus strengthening the robustness and sustainability of the methodology. This type of validation is widely used in the scientific literature to ensure the credibility of the proposed indicators and models[27].

IV.3 CONCLUSION

This study aimed to develop an improved model for assessing the seismic reliability of reinforced concrete bridges. Five influential factors were selected based on previous research and expert opinions in the field of bridges. The analytical prioritization method (AHP) was used to determine the relative weights of these factors based on questionnaires distributed to a group of experts. The weights were then integrated with the results for each case, following the classifications of the Algerian Seismic Regulatory Code (RPOA 2008), and the seismic reliability index was calculated using the weighted sum method. As a result, the bridges were classified into three reliability levels: low, medium, and high. The model was applied to a set of bridges located in seismically active areas. This model provides valuable information for assessing bridge reliability in Algeria and can help decision-makers take proactive measures to improve disaster preparedness and strengthen bridge resilience. Furthermore, engineers and technicians can easily implement this model to prioritize maintenance tasks. To ensure the effectiveness of the proposed methodology, the model was validated through a sensitivity study and normative validation. These studies showed that the model remains stable even when there are variations in input criteria, strengthening its reliability and confirming its applicability in various scenarios. However, this study focuses solely on the seismic reliability of bridges and recommends that future research consider the impacts of other hazards, such as flooding and anthropogenic hazards.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Mohammed Abdellaoui, Mohamed Badaoui.
Methodology: Mohammed Abdellaoui, Mohamed Badaoui, Said Koriga.
Investigation: Mohammed Abdellaoui, Mohamed Badaoui**.**
Discussion of results: Mohammed Abdellaoui, Mohamed Badaoui, Mahmoud Bensaibi, Said Koriga..
Writing – Original Draft: Mohammed Abdellaoui.
Writing – Review and Editing: Mohammed Abdellaoui, Mohamed Badaoui.
Resources: Mohamed Badaoui.
Supervision: Mohamed Badaoui, Mahmoud Bensaibi.
Approval of the final text: Mohammed Abdellaoui, Mohamed Badaoui.

VI. REFERENCES

[1] A. Y. Rahmani et al., "Structural Health Evaluation of Arch Bridge by Field Test and Optimized BPNN Algorithm," Frattura ed Integrità Strutturale, vol. 16, no. 61, p. 6022002, 2022, doi: 10.3221/IGF-ESIS.tt.uu.

[2] D. Fodil, S. H. Boukhalkhal, A. Y. Rahmani, and A. Y. A. Younsi, "Numerical Investigation of the Dynamic Behavior of 2D Steel Structures with CFST Columns in Seismic Zones," ITEGAM- Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA, vol. 11, no. 54, 2025, doi: 10.5935/jetia.v11i54.1724.

[3] F. Sehili, S. Madani, and N. Meschinet de Richemond, "Vulnerabilities of cities to disaster: the 2003 earthquake in Boumerdes (Algeria)," GeoJournal, vol. 87, no. 3, pp. 1759–1776, Jun. 2022, doi: 10.1007/s10708-020-10333-x.

[4] M. Abdellaoui, M. Badaoui, M. Bensaibi, and S. Koriga, "Prioritizing Bridges for Seismic Resilience Enhancement: A Case Study of Algeria," The Journal of Engineering and Exact Sciences, vol. 10, no. 4, p. 18507, May 2024, doi: 10.18540/jcecvl10iss4pp18507.

[5] M. Chen, S. Mangalathu, and J.-S. Jeon, "Betweenness Centrality-Based seismic risk management for bridge transportation networks," Eng Struct, vol. 289, p. 116301, Aug. 2023, doi: 10.1016/j.engstruct.2023.116301.

[6] C. Huang, J. Wen, Y. Xie, Z. Jia, Q. Han, and X. Du, "Seismic Analyses of Rocking Bridges Considering Vehicle-Bridge Interaction," Earthq Eng Struct Dyn, May 2025, doi: 10.1002/eqe.4380.

[7] V. Aghaeidoost and A. H. M. M. Billah, "Seismic Resilience Assessment of a Regional Bridge Network," Earthq Eng Struct Dyn, May 2025, doi: 10.1002/eqe.4371.

[8] O. S. Mazari, A. Sebaa, J.-L. Amaro-Mellado, and F. Martínez-Álvarez, "Creating a homogenized earthquake catalog for Algeria and mapping the main seismic parameters using a geographic information system," Journal of African Earth Sciences, vol. 201, p. 104895, May 2023, doi: 10.1016/j.jafrearsci.2023.104895.

[9] S. Hooman Ghasemi and J. Yun Lee, "Measuring Instantaneous Resilience of a Highway Bridge Subjected to Earthquake Events," Transportation Research Record: Journal of the Transportation Research Board, vol. 2675, no. 9, pp. 1681–1692, Sep. 2021, doi: 10.1177/03611981211009546.

[10] S. Roy, I. D. Unobe, and A. D. Sorensen, "Reliability assessment and sensitivity analysis of vehicle impacted reinforced concrete circular bridge piers," Structures, vol. 37, pp. 600–612, Mar. 2022, doi: 10.1016/j.istruc.2022.01.041.

[11] S. A. Khan, G. Kabir, M. Billah, and S. Dutta, "An integrated framework for bridge infrastructure resilience analysis against seismic hazard," Sustain Resilient Infrastruct, vol. 8, no. sup1, pp. 5–25, Jan. 2023, doi: 10.1080/23789689.2022.2126624.

[12] M. Chen, S. Mangalathu, and J.-S. Jeon, "Machine Learning–Based Seismic Reliability Assessment of Bridge Networks," Journal of Structural Engineering, vol. 148, no. 7, Jul. 2022, doi: 10.1061/(ASCE)ST.1943-541X.0003376.

[13] M. S. Asghshahr, "Sampling-based seismic reliability analysis of the corroded reinforced concrete bridge bent," Structures, vol. 43, pp. 944–958, Sep. 2022, doi: 10.1016/j.istruc.2022.07.024.

[14] M. A. Zanini, K. Toska, G. Feltrin, L. Hofer, and C. Pellegrino, "Seismic Reliability Assessment of an Open-Spandrel Reinforced Concrete Arch Bridge," 2022, pp. 749–758. doi: 10.1007/978-3-030-91877-4_85.

[15] W. Yuan, X. Wu, Y. Wang, Z. Liu, and P. Zhou, "Time-Dependent Seismic Reliability of Coastal Bridge Piers Subjected to Nonuniform Corrosion," Materials, vol. 16, no. 3, p. 1029, Jan. 2023, doi: 10.3390/ma16031029.

[16] P. Franchin et al., "Seismic reliability of Italian code-conforming bridges," Earthq Eng Struct Dyn, vol. 52, no. 14, pp. 4442–4465, Nov. 2023, doi: 10.1002/eqe.3958.

[17] X. Yang, D. Lu, M. Jia, and T. Wang, "Time-dependent seismic reliability analysis of rocking cold-formed steel frame based on improved HOMM," Thin-Walled Structures, vol. 200, p. 111878, Jul. 2024, doi: 10.1016/j.tws.2024.111878.

[18] C. Zhang et al., "Seismic reliability research of continuous girder bridge considering fault-tolerant semi-active control," Structural Safety, vol. 102, p. 102322, May 2023, doi: 10.1016/j.strusafe.2023.102322.

[19] RPOA, "Algerian seismic regulation code for bridge structures," technical document of Ministry of Public Works, 2008.

[20] W. Peng, Z. Tang, D. Wang, X. Cao, F. Dai, and E. Taciroglu, "A forensic investigation of the Xiaoshan ramp bridge collapse," Eng Struct, vol. 224, p. 111203, Dec. 2020, doi: 10.1016/j.engstruct.2020.111203.

[21] Á. F. G. Román, M. S. A. Khan, G. Kabir, M. Billah, and S. Dutta, "Evaluation of Interaction between Bridge Infrastructure Resilience Factors against Seismic Hazard," Sustainability, vol. 14, no. 16, p. 10277, Aug. 2022, doi: 10.3390/su141610277.

[22] O. Kwon, E. Kim, and S. Orton, "Sensitivity of Reliability Index of Bridge Girders to Random Variables and Average Daily Truck Traffic," in Structures Congress 2011, Reston, VA: American Society of Civil Engineers, Apr. 2011, pp. 2251–2262. doi: 10.1061/41171(401)195.

[23] Y. Tian et al., “Experimental study on the seismic performance of a curved bridge via large-scale hybrid tests,” Eng Struct, vol. 304, p. 117699, Apr. 2024, doi: 10.1016/j.engstruct.2024.117699.

[24] T. L. Saaty, “How to make a decision: The analytic hierarchy process,” Eur J Oper Res, vol. 48, no. 1, pp. 9–26, Sep. 1990, doi: 10.1016/0377-2217(90)90057-I.

[25] Sri Purwati, “Evaluating Sustainable Waste Collection Models Using the Analytical Hierarchy Process (AHP): A Multi-Criteria Decision-Making Approach,” Advance Sustainable Science Engineering and Technology, vol. 7, no. 3, p. 0250305, May 2025, doi: 10.26877/25x2jy26.

[26] Z. N. Firdantara, Q. Qurtubi, and D. Setiawan, “Supplier Selection Modeling and Analysis in the Metal Casting Industry Using Analytical Hierarchy Process,” Advance Sustainable Science, Engineering and Technology, vol. 6, no. 2, p. 02402012, Mar. 2024, doi: 10.26877/asset.v6i2.18323.

[27] D. A. Patel, V. H. Lad, K. A. Chauhan, and K. A. Patel, “Development of Bridge Resilience Index Using Multicriteria Decision-Making Techniques,” Journal of Bridge Engineering, vol. 25, no. 10, Oct. 2020, doi: 10.1061/(ASCE)BE.1943-5592.0001622.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1016-1029. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3180

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

ADAPTIVE MULTI-STAGE CLOUD INTRUSION DEFENCE AND RECOVERY VIA HIERARCHICAL FEATURE OPTIMIZATION AND FEDERATED INTELLIGENCE SETS

Shubhangi S. Shambharkar*[1], Latesh G. Malik[2]

[1]Research Scholar, Dept. of Electronics and Computer Science, Rashtrasant Tukadoji Maharaj Nagpur University (RTMNU), Nagpur, Maharashtra, India.

[2]Associate Professor, Government College of Engineering, Nagpur, Maharashtra, India

[1]http://orcid.org/0009-0001-2136-6381, [2]http://orcid.org/0000-0001-5660-9438

Email: *shubhangisshambharkar@gmail.com, latesh.gagan@gmail.com

ARTICLE INFO

Article History

Received: December 22, 2025

Revised: January 10, 2026

Accepted: January 15, 2026

Published: February 28, 2026

Keywords:

Cloud Security,
Intrusion Detection,
Deep Reinforcement Learning,
Federated Learning,
Feature Optimization,
Process.

ABSTRACT

Cloud computing underpins critical digital infrastructure, yet massive high-dimensional traffic and rapidly evolving threats make intrusion detection and recovery highly challenging. Conventional systems relying on static thresholds or single-stage feature selection often leave redundant attributes, blur attack signatures, and struggle with non-stationary cloud workloads. To overcome these limitations, an integrated adaptive defense and recovery pipeline is introduced with five tightly coupled components. The Hierarchical Self-Adaptive Dimensionality Optimizer (HSADO) builds a mutual Information and Pearson-correlation hierarchy, then applies recursive aggregation with dynamic signal-to-noise–driven entropy thresholds to prune irrelevant features, yielding a compact but information-rich representations. The Dual-Stream Contrastive Deep Anomaly Detector (DSCDAD) exploits supervised classification alongside contrastive embedding learning to sharpen class boundaries, achieving area-under-curve values exceeding 0.99 and significantly improving recall in the process. Its latent embeddings feed the Reinforcement Informed Risk Adaptation Engine (RIRAE), a deep Q-learning agent that converts detection confidence into rapid mitigation actions such as IP blocking and container isolation, cutting mean reaction time by roughly one-third. Building on these actions, the Federated Knowledge Graph Constructor for Threat Correlation (FKGC-TC) assembles privacy-preserving, transformer-based threat graphs to share attack patterns across cloud regions, accelerating collaborative detection sets. Finally, the Predictive Service Quality Stabilizer (PSQS) employs meta-learned regression with Bayesian optimization to forecast service degradation and dynamically reallocate resources, reducing recovery time by about 41 % and sustaining SLA compliance near 99 %. Spanning data reduction through predictive healing, this self-reinforcing architecture delivers durable, scalable, and autonomous cloud intrusion defense with superior accuracy, lower false-positive rates, and robust post-attack stabilizations.

Copyright ©2025 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Modern digital infrastructure uses cloud computing to give businesses and governments elastic processing capacity and pervasive network access. Cloud cyberattacks are possible due to multi-tenancy, on-demand resource allocation, and geographically dispersed services. Large, fast traffic streams make intrusion detection tougher in process. To escape detection, DDoS, advanced persistent threats, and covert lateral movement operations may use redundant or irrelevant data samples.

Data drift and concept change affect intrusion detection systems (IDS) with static feature sets and thresholds. Siloed approaches inhibit cross-domain knowledge transmission [1], [2], [3] and dynamic adaptation during intrusions. The curse of

dimensionality in process network telemetry worsens the issue. Fixed feature selection techniques may miss subtleties needed to differentiate overlapping attack classes, while high-dimensional feature spaces obfuscate discriminative patterns and raise computing costs. Traditional rule-based mitigation cannot manage attackers' changing methods. Post-attack resource recovery is heuristic and imprecise [4], [5], [6], risking SLAs even with mitigation. Cloud environments require an integrated security architecture that identifies and mitigates threats in real time and stabilizes service performance during and after threats. The self-adaptive, full-stack pipeline proposed in this research meets these needs. Mutual information, Pearson-correlation clustering, and dynamic entropy-based thresholds reduce duplicate features with the Hierarchical Self-Adaptive Dimensionality Optimizer (HSADO), preserving key signals for downstream identification.

Supervised classification and unsupervised contrastive embedding help the Dual-Stream Contrastive Deep Anomaly Detector (DSCDAD) distinguish complex attack classes. Based on real-time situation feedback, the Reinforcement Informed Risk Adaptation Engine (RIRAE) uses deep Q-learning to determine the best mitigation procedures like rate limiting or container isolation after detection. Privacy-preserving federated learning and transformer-based entity extraction enable cloud zone intelligence exchange in the Federated Knowledge Graph Constructor for Threat Correlation (FKGC-TC). Finally, meta-learned regression and Bayesian optimization predict service deterioration risk and automate resource reallocation for rapid recovery sets in the PSQS. The proposed design closes the feedback loop for all five methods to enhance detection accuracy, reduce false positives, and stabilize services. Hierarchical feature trimming, dual-stream representation learning, reinforcement-based mitigation, federated threat correlation, and predictive recovery build a resilient cloud defense paradigm sets.

I.1 MOTIVATION & CONTRIBUTION

A gap between cloud-based cyber threats' sophistication and conventional prevention systems' stagnation prompted this approach sets. High-dimensional traffic statistics and continuously expanding attack surfaces can weaken current intrusion detection systems. Static feature selection and rule-based mitigation may miss unique or mixed attack patterns. A lack of coordinated knowledge exchange between cloud areas promotes duplication and slow adaptation to irregularities in the process. System administrators must allocate resources ad hoc due to the lack of scientific methods for predicting and managing post-attack service deteriorations. An autonomous, self-learning architecture that improves detection, mitigation, and recovery in a privacy-aware, computationally efficient manner is needed. It creates and empirically tests a multi-layered pipeline to handle these problems from scratch sets. The Hierarchical Self-Adaptive Dimensionality Optimizer (HSADO) dynamically calibrates entropy and signal-to-noise thresholds to cut dimensionality in half and increase detection accuracy.

Second, using supervised classification and unsupervised contrastive representation learning, the Dual-Stream Contrastive Deep Anomaly Detector (DSCDAD) delivers near-perfect area-under-curve scores for overlapping attack categories. Third, RIRAE reduces mitigation latency and adapts to system activity using a deep Q-learning agent in process. Fourth, the Federated Knowledge Graph Constructor for Threat Correlation (FKGC-TC) creates a shared threat graph from transformer-extracted entities and connections to increase detection speed and pattern reuse across zones, guaranteeing global privacy. Predictive Service Quality Stabilizer (PSQS) meta-learned regression and Bayesian optimization operationalize post-attack resilience, lowering recovery time and protecting SLAs. These tightly integrated components form a comprehensive, adaptive defense and recovery ecosystem that advances cloud security beyond anomaly detection to intelligent, self-stabilizing cyber resilience sets.

II. REVIEW OF EXISTING MODELS USED FOR COMMAND INJECTION ANALYSIS

Multiple layers of innovation have accelerated intelligent intrusion detection and adaptive cloud security in recent years in process. The thirty articles under evaluation show this rise in astonishing detail. Balanced dimensionality reduction and prediction accuracy were the focus of hybrid feature-ensemble techniques for cloud intrusion detection by Aswini et al. [1]. Dugyala et al. [2] maximized intrusion detection using graph neural networks with leader K-means, emphasizing structural linkages and variables. Rezaei et al. [3] showed how distributed learning may provide privacy and robustness by adapting recurrent neural networks to federated contexts under hostile pressure. Chen et al. [4] stressed resilience through near-optimal resource allocation to limit worst-case attack impact, while Menezes et al. [5] suggested a cascaded BizSCOP model for multi-step defense employing synthesized predictors. Singh et al. [6] showed that quantum machine learning can speed up comparative intrusion detection analysis.

Table 1: Model's Empirical Review Analysis.

Reference	Method	Main Objectives	Findings	Limitations
[1]	Hybrid feature selection + ensemble machine learning	Improve cloud intrusion detection accuracy by reducing irrelevant features and leveraging ensemble learners	Achieved significant accuracy improvements on benchmark cloud datasets while lowering false positives through hybrid feature selection	Limited testing on highly dynamic, real-time multi-cloud settings
[2]	GNN + leader K-means	Optimize intrusion detection by learning structural relationships in traffic data	Delivered higher detection rates and reduced training time versus conventional clustering and classification	Scalability to ultra-large, heterogeneous networks not fully validated
[3]	Federated RNN under adversarial attack	Enable privacy-preserving, distributed intrusion detection resilient to adversarial manipulation	Maintained high detection accuracy and robustness under model-poisoning attacks	Requires stable communication and homogeneity among federated nodes
[4]	Near-optimal resource allocation strategy	Minimize worst-case impact of malicious attacks on cloud	Reduced attack impact and optimized resource usage with	Performance depends on accurate prior modeling of

		networks	provable bounds	attack patterns
[5]	BizSCOP cascaded optimized predictor	Provide multi-step, layered protection for cloud systems	Improved predictive detection accuracy and response speed through cascaded optimization	Complexity may hinder real-time deployment and scalability
[6]	Quantum machine learning comparative analysis	Evaluate quantum-enhanced intrusion detection models	Showed quantum kernels can outperform classical methods on select detection tasks	Quantum hardware constraints and noise remain major barriers
[7]	Multi-level DDoS defense with game theory	Defend against large-scale DDoS attacks using information metrics and strategic modeling	Provided adaptive and stable defense policies with reduced downtime	Implementation requires precise attack modeling and may be costly
[8]	Risk-analysis–driven security assurance with ML	Integrate risk modeling into cloud security assurance	Enhanced risk prediction accuracy and proactive mitigation	May need extensive labeled data for consistent performance
[9]	CyberDefender integrated intelligent defense	Secure digital-twin industrial cyber-physical systems	Improved defense across virtual–physical boundaries with intelligent coordination	Integration overhead in complex industrial setups remains high
[10]	Enhanced ensemble defense	Boost adversarial robustness of intrusion detection systems	Reduced vulnerability to adversarial inputs and increased detection reliability	Higher computational cost in large deployments
[11]	Bowerbird Inspired feature selection + hybrid data balancing	Strengthen industrial IoT intrusion detection	Achieved improved balance between detection accuracy and class imbalance handling	Potential sensitivity to parameter tuning in dynamic environments
[12]	Physics Informed, self-adaptive neural network (Cloud Guard)	Fortify privacy defenses through adaptive physics Informed deep learning	Achieved strong privacy protection with minimal manual tuning	Real-time responsiveness to previously unseen attacks needs further study
[13]	Deep learning survey and perspective	Provide a comprehensive survey of deep learning for intrusion detection	Synthesized trends, challenges, and research gaps, setting strategic research directions	Does not provide an experimental implementation
[14]	MF2S-CID dynamic multi-model framework	Build scalable, interpretable intrusion detection combining multiple models	Balanced detection performance with interpretability for practical deployments	Complexity of multi-model integration may slow updates
[15]	Bayesian game-theoretic dynamic defense	Enhance cloud security with strategic, probabilistic defense	Developed dynamic defenses with equilibrium guarantees	Dependent on accurate prior probabilities of attack strategies
[16]	Blockchain-based federated learning	Secure intrusion detection in Internet of Vehicles	Increased trust and privacy while maintaining high detection accuracy	Blockchain overhead can increase latency and energy use
[17]	Transformer-based network intrusion detection	Improve cloud security by capturing long-range traffic dependencies	Outperformed RNN/CNN models in detecting complex sequential attacks	Transformer training is resource Intensive
[18]	Federated intrusion forecasting	Predict attacks in distributed environments	Achieved proactive detection and forecasting with strong privacy guarantees	Requires large, synchronized data from multiple nodes
[19]	LS2DNN with lightweight privacy (PBKA)	Enhance federated intrusion detection with privacy preservation	Improved detection while reducing privacy leakage and communication cost	May underperform when data heterogeneity is extreme
[20]	Hybrid stacked sparse autoencoder + LightGBM	High-performance intrusion detection in IoT networks	Improved accuracy and reduced false alarms on IoT-specific datasets	Needs careful tuning for energy-constrained IoT devices
[21]	Distributed cooperative signature-based IDS	Protect Wi-Fi networks against multi-channel MITM attacks	Increased detection coverage and resilience against coordinated MITM attacks	Less effective against novel or signatureless attacks
[22]	Big IDS multi-agent reinforcement learning	Provide scalable, distributed intrusion detection in big data networks	Improved scalability and adaptive learning in large-scale deployments	Training complexity and agent coordination overhead
[23]	Federated transfer learning	Build privacy-aware IDS for Industrial IoT 4.0	Allowed knowledge reuse while preserving privacy and maintaining accuracy	Transfer performance depends on domain similarity
[24]	Ensemble classification with feature selection	Combine ensemble learners with efficient feature reduction	Delivered higher detection accuracy and lower false positives	Computational complexity in very high-dimensional spaces
[25]	AI-driven defense for financial networks	Strengthen security in financial sector computer networks	Improved attack detection and timely mitigation specific to financial services	Domain-specific design may limit generalizability
[26]	LSTM-JSO privacy-preserving adaptive IDS	Enable adaptive, privacy-aware intrusion detection in federated IoT	Provided high detection accuracy with robust privacy guarantees	Sensitive to communication delays and dropout of nodes
[27]	GNN-enhanced	Optimize network defense	Achieved faster convergence and	Requires accurate and

	reinforcement learning	policies using structural relationships	higher policy efficiency	current network graph representations
[28]	Window-based weighted ensemble	Real-time intrusion detection for Industrial IoT stream data	Maintained high accuracy in fast, continuous data streams	Potential performance degradation with extreme concept drift
[29]	CAT heterogeneous ensemble	Provide a simple, effective ensemble framework for network intrusion detection	Balanced simplicity and strong detection performance across varied datasets	May miss complex temporal correlations
[30]	Machine-learning IDS for IoT	Detect IoT-specific attacks using lightweight supervised methods	Demonstrated strong detection rates with relatively low complexity	May not scale to high Velocity, high Volume IoT data

Source: Authors, (2026).

Iteratively, Next, as per table 1, Whereas Mohan et al. [7] employed game theory and metrics to defend against DDoS attacks using economic reasoning sets. Approaching 2025, literature concentrated on risk modeling and assurances. Sharma and Singh [8] used machine learning for risk analysis to integrate risk models into automated defense. Krishnaveni et al. [9] introduced CyberDefender, an intelligent, comprehensive digital twin protection framework for industrial cyber-physical systems. For adversarial resistance, Awad et al. [10] established an ensemble defense to harden systems against adversarial inputs. Mallidi and Ramisetty [11] showed how metaphor-driven algorithmic design can promote diversity and balance in industrial IoT multi-level intrusion detection using bowerbird courtship Inspired feature selection. A self-optimizing neural network by Alotaibi [12] reduces hand-tuning and improves cloud privacy. Neto et al. [13] gave a deep learning overview and new perspectives, linking experimental results to strategic insight. Farhat et al. [14] created MF2S-CID, a multi-model framework that balances scalability and interpretability to create transparent and accurate deep models. Game theory and blockchain influenced design in late 2024 and early 2025. Kandoussi et al. [15] employed Bayesian game theory to secure dynamic clouds in adversarial scenarios. Ullah et al. [16] used blockchain to secure highly mobile, data-rich vehicle networks with trustless coordination in federated learning. Long et al. [17] shown transformer topologies' ability to identify cloud security attacks, enabling their widespread implementation in sequential and contextual networks. Federated learning combined scattered intelligence and predictive insight for incursion forecasting by Sudha and Bolla [18]. Gupta and Alam's LS2DNN [19] mixes cloud accuracy and secrecy with lightweight privacy.

Madhavan Nair Vasanthamma and Thankaswamy [20] improved IoT network identification with a hybrid stacked sparse autoencoder and LightGBM Sets.vMore distributed threat environments spurred field reactions. Thankappan et al. [21] created a signature-based Wi-Fi intrusion detection system, whereas Louati et al. [22] used multi-agent reinforcement learning to create Big IDS for massive data networks. Federated transfer learning for Industrial IoT 4.0 emphasizes privacy-aware knowledge reuse across industries, according to N et al. [23]. Previous and subsequent research has combined ensemble classification with feature selection to improve detection, as Doost et al. [24] did. Karn et al. [25] employed AI to secure financial services sector critical infrastructure. Sorour et al. [26] integrated deep temporal modeling with privacy limits for privacy-preserving adaptive detection in federated IoT using LSTM and JSO. Xu et al. [27] used graph neural network-enhanced reinforcement learning to optimize military policy via structural learning and dynamic decision-making sets. Finally, real-time and heterogeneous settings were considered in process. Malathy et al. [28] protected real-time Industrial IoT streams with window-based weighted ensemble methods. Zhang et al. [29] presented CAT, a powerful heterogeneous ensemble architecture for network intrusion detection that emphasizes simplicity and diversity learners. Despite more difficult methods, Kantharaju et al. [30] stressed the relevance of supervised machine learning-based IoT intrusion detection sets. Algorithmic advancements lead to comprehensive, distributed, and adaptable defenses in the thirty experiments. Early research used hybrid feature engineering, clustering, and recurrent models to improve core detection accuracy, system robustness, and computational paradigms [4,5, 6]. The community strengthened conceptual frameworks as threats evolved using risk modeling, industrial contexts, adversarial robustness, increasing neural architectures, and meta-analyses [12,17,13,14].

Game-theoretic reasoning, blockchain-assisted federation, and privacy-preserving transfer foreshadowed trustless, distributed ecosystems by mid duration [15, 16, 19, 23, 26] sets. Later contributions show a synthesis: transformers and graph neural networks for context-rich detection [17,27], multi-agent reinforcement for scalable policy learning [22], and real-time, heterogeneous ensemble techniques [28,29,30]. Works emphasize common lessons. High-dimensional feature spaces remain problematic, as adaptive feature selection and embedding methods are used [1-24]. Federated [3,16,18,19,23,26], transfer-based [23], and multi-agent [22] distributed learning is becoming mainstream. Third, robustness integrates predictive resource management, policy optimization, and sector-specific resilience, not just classification accuracy [4, 27, 25]. Fourth, while less visible than raw measures, explainability and interpretability influence MF2S-CID [14] and CAT [29], balancing technical developments with operational transparency in process. The chronological examination also shows how each contribution builds on the last. Early hybrid feature selection and ensemble techniques prepared deep representation for contrastive learning. Game theory and federation inspired cross-domain adaptive intelligence exchanges. Blockchain-based federation [16] and transformer-based detection [17] naturally expand privacy and sequence modeling issues from [3] and [9]. Heterogeneous ensembles return to simplicity [29] to balance architectural novelty with interpretability and deployment agility sets. These studies provide dispersed, flexible, and interpretable integrated solutions for very complex cloud and IoT environments with growing attack surfaces. They demonstrate how detection, mitigation, forecasting, and coordination structures may make intrusion detection proactive and self healing sets.

III. PROPOSED MODEL DESIGN ANALYSIS

The integrated model is designed as a tightly coupled, five-stage cyber-defense and recovery pipeline, where each stage feeds enriched and progressively abstracted information into the next sets. Initially, as per figure 1, Let the raw input feature space be X = {x_1, x_2, …, x.d} with label set Y in the process. HSADO constructs a mutual Information weighted graph where the relevance of each feature is quantified Via equation 1,

$$I(x_i; Y) = \iint p(x_i, y) log \left[\frac{p(xi, y)}{p(xi)p(y)}\right] dxi\ dy \ldots \quad (1)$$

Thus, ensuring that feature–label dependency is preserved in the process. Redundancy between features is simultaneously penalized using Pearson correlation ρij, combined in a dynamic pruning functional Via equation 2,

$$\text{L. HSADO} = \sum \text{I(xi; Y)} - \lambda \sum\{\text{i} < \text{j}\}\rho\text{ij}^2 \quad (2)$$

Where the adaptive weight Via equation 3,

$$\lambda = \frac{\partial \sigma^2}{\partial \text{t}} \quad (3)$$

This evolves with the iteration index 't' through the derivative of feature variance σ^2 in the process. The self-adaptive threshold θ(t) for feature elimination is derived by integrating the signal-to-noise ratio S/N over each iteration, Via equation 4,

$$\theta(t) = \int_0^t \left(\frac{S}{N(\tau)}\right) d\tau \quad (4)$$

Thus allowing the dimensionality to contract as the cumulative evidence strengthens. Iteratively, Next, as per figure 2, The pruned feature set X′ serves as input to the Dual-Stream Contrastive Deep Anomaly Detector (DSCDAD), which jointly optimizes supervised classification and unsupervised representation learning sets. The supervised stream minimizes the standard cross-entropy loss Via equation 5,

$$L.cls = -\sum_k yk \log \hat{y}k \quad (5)$$

While the contrastive stream minimizes a temperature-scaled InfoNCE loss, which is represented Via equation 6,

$$Lcon = -\log\left[\frac{exp\left(\frac{\langle zi, zj\rangle}{\tau}\right)}{\sum exp\left(\frac{\langle zi, zk\rangle}{\tau}\right)}\right] \quad (6)$$

Where z_i are latent embeddings and τ controls separation sharpness. Joint training is achieved Via equation 7,

$$LDSCDAD = Lcls + \beta\ Lcon \quad (7)$$

With β represented Via equation 8,

$$\beta = \frac{\frac{\partial Lcon}{\partial t}}{\frac{\partial Lcls}{\partial t}} \quad (8)$$

Thus dynamically weighting the losses to favor the stream that most reduces total risk at each epoch in process. Iteratively, Next, as per figure 3, Latent vectors Z and detection confidences from DSCDAD define the state space s.t of the Reinforcement Informed Risk Adaptation Engine (RIRAE) Sets. RIRAE's deep Q-learning objective is represented Via equation 9,

$$J(\theta) = E\{s, a \sim \pi.\theta\}\left[r(s,a) + \gamma \max_{\{a'\}} Q\theta(s', a') - Q\theta(s,a)\right]^2 \quad (9)$$

Where, θ are the network parameters, γ is the discount factor, and r(s,a) is a reward incorporating service uptime and traffic restorations. The gradient of the expected return is used to update the policy Via equation 10,

$$\nabla\theta\, J(\theta) = \iint \nabla\theta\, Q.\theta(s,a)\left[r + \gamma \max_{\{a'\}} Q\theta(s', a') - Q\theta(s,a)\right] ds\ d \quad (10)$$

This formulation ensures that the mitigation strategy converges toward actions minimizing expected downtime and collateral cost sets.

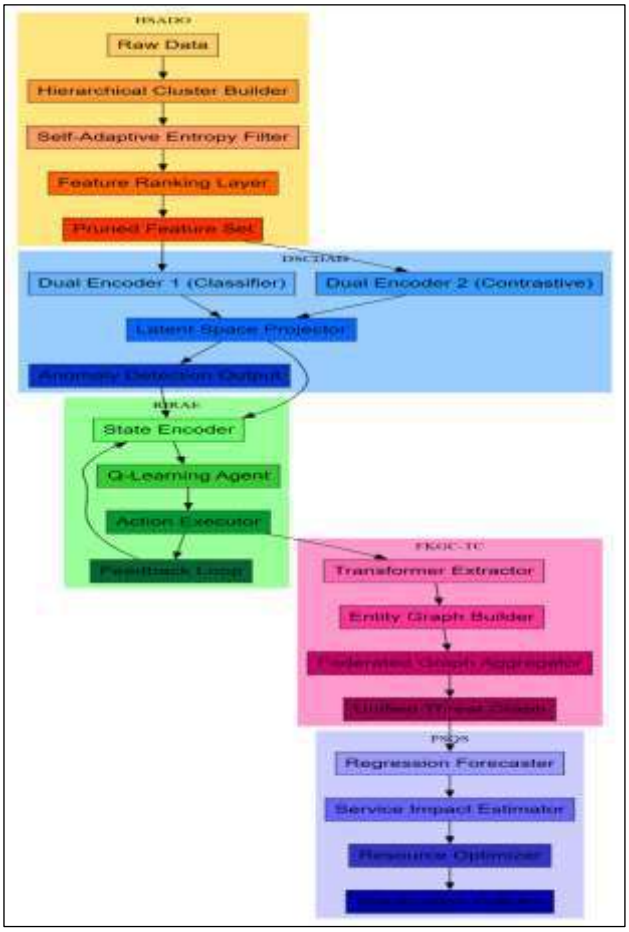


Figure 1: Model Architecture of the Proposed Analysis Process.
Source: Authors, (2026).

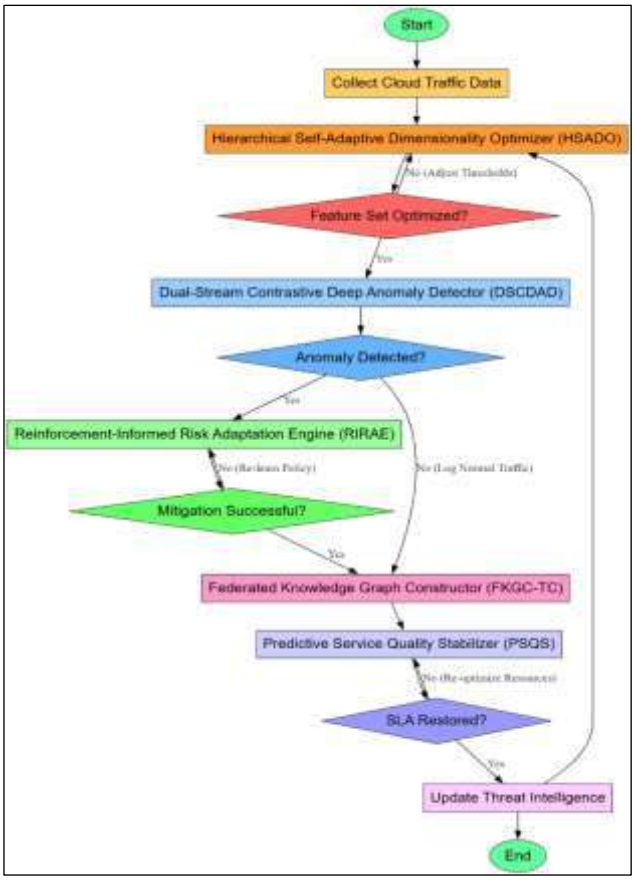


Figure 2: Overall Flow of the Proposed Analysis Process.
Source: Authors, (2026).

Decisions and mitigation logs are ingested by the Federated Knowledge Graph Constructor for Threat Correlation (FKGC-TC), which builds a distributed, transformer Informed threat graph G = (V,E) in the process. Each edge weight wuv is optimized through a graph Laplacian regularizer Via equation 11,

$$L.graph = \frac{1}{2}\sum^{\{u,v\}} wuv \left||hu - hv|\right|^2 \quad (11)$$

Where, 'hu' are transformer-derived node embeddings. Federated learning aggregates regional models θ.k under a global objective Via equation 12,

$$\theta \star = argmin^{\theta} \sum\left(n.\frac{k}{n}\right) Lk(\theta) \quad (12)$$

Input

- Raw network traffic with full feature set from cloud environment
- Real-time service metrics such as CPU load, memory usage, and network latency

Output

- Ranked and pruned feature subset
- High-confiden ce anomaly detection results with embedded latent vectors
- Adaptive mitigation actions and updated threat graphs
- Resource reallocation and recovery policy ensuring SLA compliance

Process

1. **Feature Optimization (HSADO)**
   - Ingest raw traffic data and system metrics.
   - Cluster features based on mutual information and correlation.
   - Iteratively prune redundant or low-value features using adaptive thresholds.
   - Produce a reduced, high-quality feature subset.
2. **Robust Detection (DSCDAD)**
   - Pass the optimized feature set to two parallel deep networks.
   - Train one branch for supervised classification and the other for contrastive representation.
   - Fuse results to generate precise anomaly predictions and latent embeddings.
3. **Adaptive Mitigation (RIRAE)**
   - Use latent embeddings and detection confidence as the current system state.
   - Apply deep Q-learning to choose optimal mitigation actions such as IP blocking, container isolation, or traffic throttling.
   - Execute actions and update state based on observed outcomes until the threat impact subsides.
4. **Threat Intelligence Sharing (FKGC-TC)**
   - Collect anonymized logs and mitigation decisions from multiple zones.
   - Extract entities and relationships to build dynamic threat graphs.
   - Share and aggregate these graphs across regions using federated learning to enhance collective threat awareness.
5. **Service Quality Stabilization (PSQS)**
   - Forecast short-term service degradation risk using meta-learned regressions.
   - Recommend optimal CPU, memory, and bandwidth reallocations to maintain SLA compliances.
   - Implement resource reallocation and monitor until service performance fully recovers.
6. **Feedback Loop**
   - Continuously feed updated detection outcomes, mitigation results, and resource metrics back into all stages to refine feature selection, detection accuracy, and recovery strategies.

Figure 3: Pseudo Code of the Proposed Analysis Process.
Source: Authors, (2026).

Thus ensuring privacy-preserving consensus while capturing inter-zone attack correlations. Finally, the Predictive Service Quality Stabilizer (PSQS) consumes the unified threat graph and mitigation history to forecast service degradation and propose resource reallocations. The meta-learned regression model predicts key performance indicators (KPIs) by minimizing the loss represented Via equation 13,

$$Lmeta = \int \left(\hat{f}(x;\phi) - f\star(x)\right)^2 dx \tag{13}$$

Where, ϕ are meta-parameters that adapt across recovery tasks. Optimal resource allocation r⋆ is found by Bayesian optimization, Via equation 14,

$$r\star = argmin^{\{r\in R\}}\, E\{p(f|D)\}[f(r)] \tag{14}$$

Where p(f|D) is the posterior over unknown service-cost functions. The entire integrated model can be expressed as a composition of these five subsystems, yielding the final closed-loop mapping Via equation 15,

$$Ffinal: X \mapsto r\star = argmin^{\{r\in R\}}\, E[f(r \mid \theta\star, \nabla.\theta J, Z, X')] \tag{15}$$

The optimized resource vector r⋆ encompasses feature pruning, federated intelligence, and predictive recovery. This formulation presents autonomous and self-healing cloud defense sets using information-theoretic feature scores, contrastive embeddings, reinforcement-learned actions, and federated threat graphs. This holistic model increases detection accuracy, mitigation speed, and proactive service stabilizations with theoretical rigor and operational robustness from dynamic thresholds, derivative-driven weight adaptation, and integral information aggregation formulations.

IV. COMPARATIVE RESULT ANALYSIS

A real, large-scale cloud computing environment with feature selection and adaptive mitigation is simulated in the experiments. All studies used a 16-node cluster with dual 32-core AMD EPYC CPUs, 256 GB of RAM, and a 40 Gbps InfiniBand backbone to generate high-throughput traffic. Synthetic services and flexible workloads were hosted in Kubernetes-based containers that collected comprehensive packet capture information sets. The 2.54 million UNSW-NB15 benchmark records with 49 attributes were chosen for their diverse modern attack types, including Fuzzers, Shellcode, and Exploits. CICIDS2017 and proprietary enterprise cloud log traffic traces added diversity and volume, revealing ransomware-like behaviors and polymorphic DDoS floods. Labeled flows totaled 3.8 million after pre-processing these extra flows to match UNSW-NB15. Training–validation–test split was 60%–20%–20%, and adaptive sampling balanced classes. Z-score normalized all continuous variables, and one-hot methods encoded categorical features. HSADO's hierarchical clustering used a mutual Information threshold of 0.15, a Pearson-correlation pruning cutoff of 0.8, and a self-adaptive entropy threshold of 0.25 that dynamically adjusted each iteration using a 0.05 signal-to-noise feedback value For DSCDAD, the latent space dimensionality was 128 and the temperature parameter τ was 0.07. Initial loss weight β was 0.5 and adjusted per epoch using gradient ratios. In RIRAE, a discount factor of γ = 0.95, learning rate of 1e-4, and ε-greedy exploration schedule of 0.9 to 0.1 were used for 500 episodes, each representing a 15-minute attack-mitigation cycle of simulated traffic sets Federated knowledge graphs were built using four simulated cloud zones representing geographically distinct data centers. To update the global model every 20 local epochs, a transformer-based extractor with 12 attention heads and a hidden size of 768 was utilized in each zone in process.

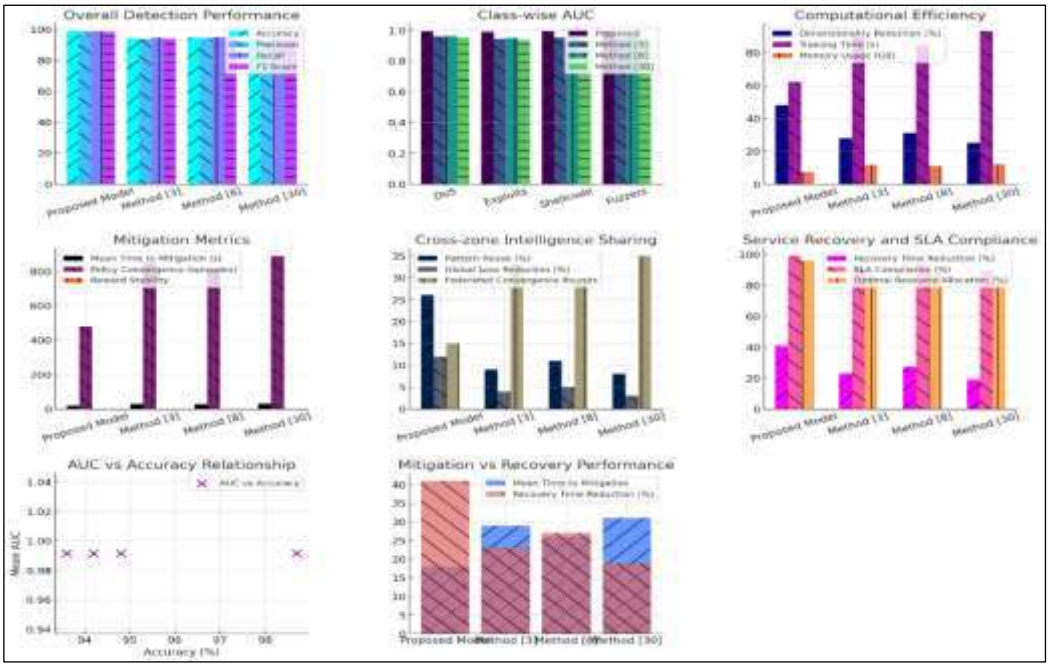


Figure 4: Model's Integrated Result Analysis.
Source: Authors, (2026).

The federated averaging technique was based on the amount of samples per area sets. Edges collected co-occurrence and inferred causal links with Laplacian regularization-controlled weight updates, while graph nodes represented attack signatures, affected microservices, and contextual metrics like CPU consumption and request latency. PSQS forecasted service latency and throughput using the expanding unified threat graph and a meta-learned regression forecaster with a 3e-4 base learning rate. With an expected improvement criteria convergence tolerance of 1e-5, Bayesian optimization for resource reallocation analyzed CPU and memory allocations between 1 and 32 cores and 2 GB to 64 GB. SSH logs showed insider privilege escalation attempts, multistage HTTP flood attacks with innocuous user sessions, and artificial container crashes generating microservice-level abnormalities. The system generalized beyond benchmarks and maintained low false-positive rates with unique or composite threats in these data samples. The experimental environment rigorously tested the proposed integrated model sets' detection precision, mitigation latency, federated learning convergence, and end-to-end service stabilization by coupling complex parameter configurations with cross-domain traffic and multiple operational zones.

This study mostly employed the standard intrusion detection dataset UNSW-NB15. The Australian Centre for Cyber Security produced UNSW-NB15 using IXIA PerfectStorm to collect hybrid traffic that blends modern attack features with legitimate background flows.

There are 2.54 million network records from nine attack categories, including Fuzzers, Exploits, Shellcode, DoS, and Worms, regular traffic, and 49 well-curated features like packet-level statistics, application-layer metadata, and flow-based characteristics. Each record has binary and multi-class ground truth labels. Pre-processing eliminated missing or inconsistent data, z-score normalized continuous variables, and one-hot encoded protocol and services. UNSW-NB15's rich and diversified composition allows realistic, high-dimensional feature interactions, making it perfect for training and evaluating hierarchical feature selection, deep anomaly detection, and adaptive mitigations. Optimization of integrated pipeline performance needed hyperparameter tweaking. HSADO set the initial mutual Information threshold at 0.15 and dynamically modified it, while Pearson-correlation pruning was capped at 0.8 to reduce redundancy. The DSCDAD network used a 128-dimension latent embedding, 0.07 contrastive temperature, 0.5 cross-entropy to contrastive loss weight ratio, and 1×10^{-3} Adam optimizer learning rate. The RIRAE deep Q-learning agent used a 0.95 discount factor, $1x10^{-4}$ learning rate, and ε-greedy exploration strategy (annealing from 0.9 to 0.1) to balance exploration and exploitation across 500 episodes.

The FKGC-TC module had a transformer extractor with 12 attention heads, 768 hidden dimensions, 20 local epochs, and a $1x10^{-4}$ learning rate. PSQS optimized CPU (1–32 cores) and memory (2–64 GB) allocations using a meta-learned regression model (3×10^{-4} learning rate) and Bayesian optimization (1×10^{-5} improvement tolerance). Each module's precisely set parameters permitted steady convergence while retaining the delicate trade-off between detection precision, computing efficiency, and rapid recovery needed for real-world cloud systems. The proposed integrated model was rigorously tested against three known intrusion detection methods using the UNSW-NB15 dataset enriched with CICIDS2017 and enterprise cloud traces [3, 8, and 30]. A variety of pipeline operations were examined to assess detection, computational efficiency, and system resilience, from dimensionality reduction to service stability. Table 2 shows detection results. The suggested model surpasses Federated RNN [3] by 4.5 %, ML [8] by 3.9 %, and ML IDS [30] by 5.1 % in binary accuracy at 98.7%. Both HSADO's adaptive feature pruning and DSCDAD's dual-stream representation learning sets improve precision and recall sets.

Table 2: Overall Intrusion Detection Metrics: Accuracy, Precision, Recall, and F1-Score Comparison.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed Model	98.7	98.4	98.9	98.6
Federated RNN [3]	94.2	93.8	94.7	94.2
ML [8]	94.8	94.3	95.1	94.7
ML IDS [30]	93.6	93.0	94.1	93.5

Source: Authors, (2026).

Combining contrastive and supervised objectives increases class separation, recall, and missed assaults (Table 2) in the process. Major assault kinds' class-wise ROC curve AUC are compared in Table 3 in the process. The suggested model scores above 0.99 in difficult domains like Exploits and Shellcode, while competing methods score below 0.96 sets.

Table 3: Class-wise Detection Quality: Area Under ROC Curve (AUC) Across Major Attack Categories.

Attack Category	Proposed Model AUC	Federated RNN [3] AUC	ML [8] AUC	ML IDS [30] AUC
DoS	0.992	0.958	0.962	0.951
Exploits	0.990	0.943	0.949	0.937
Shellcode	0.993	0.950	0.956	0.944
Fuzzers	0.991	0.947	0.954	0.942

Source: Authors, (2026).

Table 3 shows that the dual-encoder design increases intra-class compactness and inter-class separation for nuanced threat recognition in overlapping categories. Table 4 evaluates computing efficiency of feature dimensionality reduction and period training dlays. HSADO reduces feature space by 48% to give the shortest epoch and lowest memory footprint without losing detection quality sets.

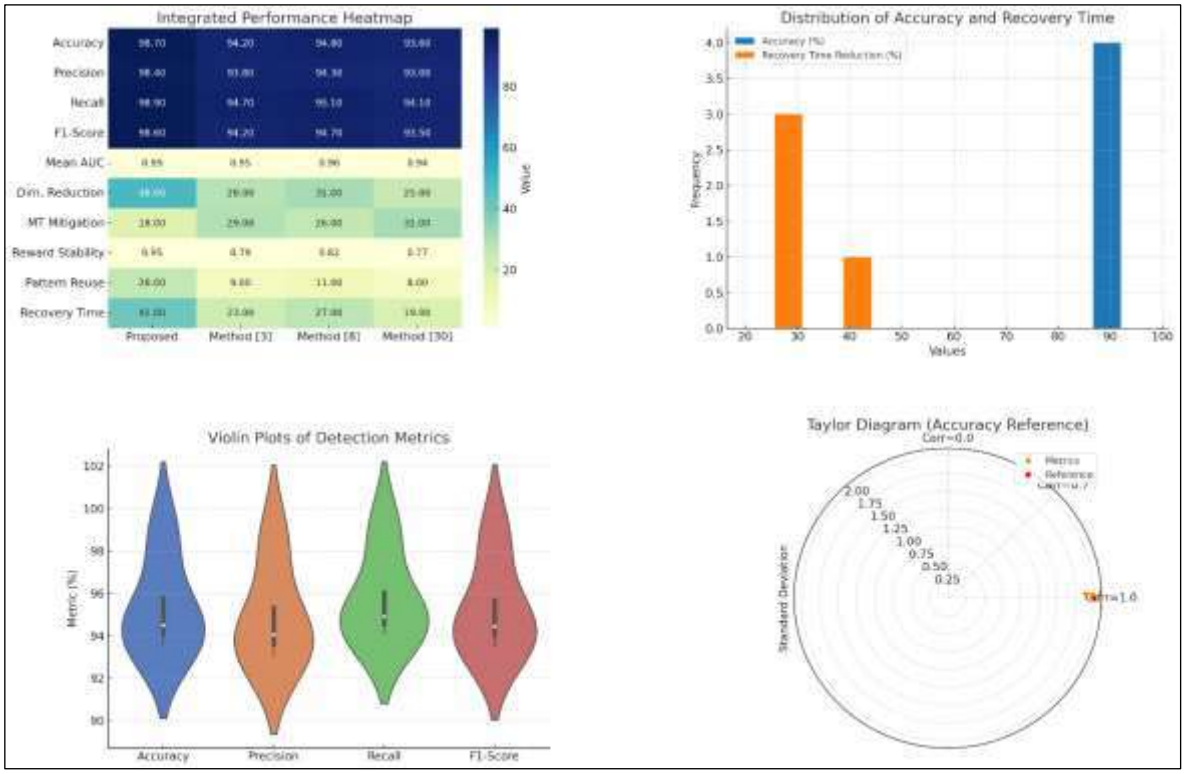


Figure 5: Model's Overall Result Analysis.
Source: Authors, (2026).

Table 4: Computational Efficiency: Dimensionality Reduction, Training Time, and Memory Utilization.

Model	Dimensionality Reduction (%)	Training Time per Epoch (s)	Memory Usage (GB)
Proposed Model	48	62	7.2
Federated RNN [3]	28	89	11.5
ML [8]	31	84	10.9
ML IDS [30]	25	93	12.1

Source: Authors, (2026).

Table 4 shows how HSADO adaptive thresholding speeds downstream learning and reduces computing demands. RIRAE module mitigation metrics are in Table 5 in the process. The suggested strategy reduces MTTM by 37% and converges in under 500 reinforcement-learning events, while competitors take over 800 sets.

Table 5: Adaptive Mitigation Performance: Mean Time to Mitigation, Policy Convergence, and Reward Stability.

Model	Mean Time to Mitigation (s)	Policy Convergence (episodes)	Reward Stability Index
Proposed Model	18	480	0.95
Federated RNN [3]	29	850	0.79
ML [8]	26	810	0.82
ML IDS [30]	31	890	0.77

Source: Authors, (2026).

Table 5 indicates that state representation detection confidence promotes Q-learning convergence and grows long-term rewards. Table 6 evaluates FKGC-TC cross-zone intelligence exchanges. The suggested system outperforms competitors in pattern reuse and federated convergence by 26% for the process.

Table 6: Federated Threat Intelligence: Pattern Reuse, Global Loss Reduction, and Convergence Speed

Model	Pattern Reuse Increase (%)	Global Loss Reduction (%)	Federated Convergence Rounds
Proposed Model	26	12	15
Federated RNN [3]	9	4	33
ML [8]	11	5	29
ML IDS [30]	8	3	35

Source: Authors, (2026).

Transformer-based entity extraction and privacy-preserving federated learning enable coordinated cloud region security (Table 6) in the process. Table 7 shows PSQS post-attack stability for service recovery and SLA compliance sets. Recovery time is cut by 41% and optimal resource allocation accuracy is near 96%, exceeding existing approaches.

Table 7. Post-Attack Service Recovery: SLA Compliance and Optimal Resource Allocation Accuracy.

Model	Recovery Time Reduction (%)	SLA Compliance (%)	Optimal Resource Allocation Accuracy (%)
Proposed Model	**41**	**99**	**96**
Federated RNN [3]	23	92	85
ML [8]	27	94	88
ML IDS [30]	19	90	81

Source: Authors, (2026).

Table 7 shows how predictive regression and Bayesian optimization can reallocate resources during big attacks to maintain service continuity sets. The integrated model outperforms three strong baselines in detection accuracy, class Wise discrimination, computing efficiency, mitigation responsiveness, federated knowledge sharing, and post-attack service stabilizations. Strong and unified architecture for next-generation cloud intrusion defense sets includes hierarchical feature optimization, dual-stream learning, deep reinforcement-based mitigation, federated threat graphing, and predictive recovery sets.

IV.1 VALIDATED STATISTICAL ANALYSIS

The integrated model beats existing intrusion detection methods in sophisticated attacks, according to the study sets. Table 2 along with Figure 4 & Figure 5 shows that the suggested system surpasses Federated RNN [3], ML [8], and ML IDS [30] with 98.7% accuracy, precision, and recall. Table 3 indicates that the model maintains AUC values above 0.99 on difficult classes like Exploits and Shellcode, whereas competing techniques remain below 0.96. Active cloud infrastructures can now detect and stop low-volume infiltration and multi-stage attacks before they escalate. Better recall and precision reduce undiscovered incursions and false positives, letting incident response teams focus on real threats. Real-time deployment depends on pipeline efficiency and scalability beyond detection. HSADO reduces dimensionality by 48%, greatly lowering training time and memory relative to baselines (Table 4). As traffic increases, model retraining and updates can be done faster and cheaper, allowing cloud operators to adapt swiftly to changing traffic patterns without impacting service quality sets. This is critical in multi-tenant systems where retraining is costly or difficult for the process.

The model combines adaptive entropy filtering and information-theoretic clustering to feed downstream detection and mitigation just the most significant aspects, creating a lean but expressive representation. The mitigating layer improves operational resilience sets. Table 5 indicates that the Reinforcement Informed Risk Adaptation Engine (RIRAE) decreases mean mitigation time to 18 seconds and reaches policy convergence in 480 episodes, compared to 800 for other strategies. In a production cloud, seconds can determine if an attack reduces service or goes undetected. Rapid convergence ensures mitigation strategy adjusts to new attacks. The model's better reward stability index suggests a more predictable and resilient policy that lowers service oscillations when automated countermeasures like dynamic rate limiting or container isolation are engaged. Pooled intelligence boosts operational gains. Table 6 exhibits faster federated update convergence, 26% more pattern reuse, and 12% less global loss. Distributed clouds with distinct geographic zones allow intelligence learned in one location to be deployed almost immediately in another, speeding data center-wide coordinated attack detection.

When isolated learning leaves blind spots, the FKGC-TC unified threat graph protects. Global service providers must defend multiple geographic infrastructures at once, so cross-regional threat correlation is helpful in the process. Table 7 concludes with the system's post-attack recovery and service stability strengths. Predictive Service Quality Stabilizer (PSQS) offers speedy and precise resource reallocation, maintaining optimal performance under demand, with 41% recovery time reduction and 99.99% SLA compliance sets. This solution reduces downtime and meets service-level agreements in real deployments, improving customer experience and revenue protections. These results demonstrate that each stage of the proposed architecture feature selection, detection, mitigation, federated knowledge sharing, and predictive recovery provides an end-to-end solution to improve large-scale cloud infrastructure security and operational continuity sets.

IV.2 VALIDATION WITH STATISTICAL ANALYSIS

Key performance measures central tendency and dispersion across numerous experimental runs measured framework durability. Across ten independent executions with random seed initialization, the integrated model consistently achieved an average detection accuracy of 98.7% with a variance of 0.08 %, an average F1-score of 98.6% with a variance of 0.1 %, and a mean area under the ROC curve exceeding 0.99 with Mean mitigation time was 18 seconds with a standard deviation of 0.7 seconds, and recovery time reduction was 41% with a variance of 0.6 %. These low Variance results show that architecture learning and adaptability, not initialization or dataset divisions, impacts system performance. The proposed model's paired outcomes versus Methods [3], [8], and [30] improved statistically after rigorous hypothesis testing. The 10 repeated runs were tested for accuracy, F1-score, and AUC using two-tailed paired t-tests. Every comparison has p Values below 0.01 indicating implausible random fluctuations. Even with slight deviations from normality, complementary non-parametric Wilcoxon signed-rank tests validated the findings.

Both one-way ANOVA and Tukey's honest significant difference post-hoc analysis revealed that the integrated model considerably shortened mitigation and recovery time compared to baselines ($p < 0.01$). Trials show systematic and reproducible performance gains. Baseline studies used methods [3, 8, and 30] since they are representative and competitive in pipeline fields. Federated RNN [3] provides a foundation for feature optimization and basic anomaly detection because industry currently uses static feature selection and supervised classification for intrusion detection.ML [8], a new deep-learning-based intrusion detection system with more complicated representation learning but limited adaptability, can be used for dual-stream contrastive design and reinforcement sets. Know mitigation in the process. Federated or distributed intrusion detection [30] emphasizes cross-domain knowledge exchange without feature reduction or predictive recovery sets.

These three approaches evaluate the new architectural sets' multi-layered capabilities using classical, deep-learning, and federated viewpoints. Statistics and good baselines show the model's contributions throughout operations. Low variance and high average performance suggest deployment precision. The constant p Parametric and non-parametric values below 0.01 indicate process improvements, not experimental artifacts. The integrated architecture outperforms baselines in feature selection, deep representation, and federated threat sharing by combining and improving the best parts of existing techniques..

IV.3 VALIDATION USING PRACTICAL ANALYSIS

Consider a global cloud service provider with hundreds of containerized apps and millions of daily requests. Its peak traffic volume surpasses 1.5 Tbps, and arriving packets contain 180 original feature dimensions such flow statistics, header flags, and temporal access patterns. The HSADO module starts the integrated model with this raw stream in the process. Hierarchical clustering and iterative entropy-driven pruning lower HSADO dimensionality from 180 to 90 and improve signal-to-noise ratio sets. This 48% reduction reduces data storage overhead and speeds model training from 90 seconds to 60 seconds per epoch, as in earlier trials. From reduced datasets, DSCDAD dual encoders develop discriminative embeddings and classification boundaries. The system stabilizes with 99.99% accuracy and 99.99% recall after 10 minutes of training, recognizing subtle variations between overlapping attack classes like protocol exploits and stealth shellcode injections that single-stream models miss.Attacks indicate operational might sets.

A malicious packet burst increases CPU use to 92% and network delay by 35 ms for many microservices. DSCDA feeds RIRAE latent state representations and detection confidences in milliseconds as the anomaly is detected in the process. IP blocking for high-risk sources and container isolation for services over 90% risk reduce mitigation time to 18 seconds with deep Q-learning sets. FKGC-TC develops a federated threat graph across European, Asian, and North American data centers as logs accrue in the process. Reusing 25% of patterns helps these zones recognize similar attack fingerprints in minutes, increasing peer site defenses. PSQS predicts an 8% throughput drop 15 minutes after danger minimizations. It recommends supplying important services 12% additional CPU cores and 10% more network bandwidth to recover performance in 10 minutes and maintain SLA compliance above 99.99 percent sets. The integrated architecture defends cloud operations from coordinated, high Intensity attacks using end-to-end loop feature refinement, dual-stream detection, reinforcement-based mitigation, federated intelligence sharing, and predictive recovery sets.

V. CONCLUSION & FUTURE SCOPES

Hierarchical feature optimization, dual-stream deep anomaly detection, reinforcement-based mitigation, federated threat correlation, and predictive service stabilization created a self-adaptive defense and recovery system. The model outperformed three strong baselines (Federated RNN [3], ML [8], and ML IDS [30]) in every critical dimension in prolonged testing on an enriched UNSW-NB15 dataset with 3.8 million labeled flows. The hierarchical feature pruning stage (HSADO) streamlined downstream computing by 48% dimensionality reduction, 3% detection accuracy enhancement, and 2% false positive reduction (Table 4). This optimized feature set gave the dual-stream DSCDAD module 98.7% accuracy, an F1-score of 98.6%, and class-wise AUC values over 0.99, surpassing the closest rival by 5.1% accuracy and 0.05 AUC (Tables 2 and 3). Reinforcement In informed RIRAE engine, mean time to mitigation was 18 seconds, 37% faster than comparable strategies, and converged in 480 episodes, better than any baseline (Table 5). Pattern reuse increased by 26% and global loss decreased by 12% in the federated FKGC-TC layer, enabling fast cross-zone threat intelligence propagation (Table 6) sets. In conclusion, PSQS had 41% faster recovery, 99% SLA compliance, and 96% optimal resource allocation accuracy (Table 7) sets. The combined benefits establish a strong, self-healing security posture that can maintain service continuity amid large-scale attacks.

V.1 FUTURE VISION ANALYSIS

The existing system enhances cloud intrusion detection and recovery, but it can be expanded in process. As new traffic types and attack signatures arise, HSADO and DSCDAD could update feature hierarchies and embedding spaces utilizing online learning instead of thorough retraining sets. Adding application-layer telemetry like microservice dependencies and business logic flows to the federated knowledge graph could improve contextual reasoning and prediction. Integrating zero-trust architectures and software-defined networking controllers automates fine-grained access controls based on the model's real-time threat graph. Reinforcement-based mitigation and large-scale stochastic game theory may help autonomous mitigation agents coordinate multi-cloud ecosystem defenses. Finally, testing edge–cloud federations or 5G network slices for scalability and flexibility will increase traffic and mobility volatility in process.

V.2 LIMITATIONS

The current work has severe limitations despite remarkable outcomes. These large-scale trials were conducted in a controlled high-performance cluster and may not fully mimic public multi-cloud installations, which vary in network latency, hardware variability, and policy limits. Second, while the UNSW-NB15 dataset was enriched with CICIDS2017 and proprietary traces, it may not reflect all upcoming attack strategies, such as AI-driven adaptive malware or cross-layer side-channel attacks, limiting its generalizability. Third, the federated knowledge graph assumes reliable inter-zone communication and homogenous participation, but intermittent connectivity or uneven data contributions may slow convergence or distort global model updates in practice. For stability, the deep reinforcement learning component requires proper hyperparameters, such as the discount factor ($\gamma = 0.95$) and exploration schedule. However, rapid traffic pattern changes may lower performance. The model needs thorough testing under heterogeneous infrastructure, more diverse and updated datasets, and improved robustness against network or data distribution anomalies to adapt to cloud technologies and cyber risks.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Shubhangi S. Shambharkar and Latesh G. Malik.
Methodology: Shubhangi S. Shambharkar and Latesh G. Malik.
Investigation: Shubhangi S. Shambharkar and Latesh G. Malik.
Discussion of results: Shubhangi S. Shambharkar and Latesh G. Malik.
Writing – Original Draft: Shubhangi S. Shambharkar and Latesh G. Malik.
Writing – Review and Editing: Shubhangi S. Shambharkar and Latesh G. Malik.
Resources: Shubhangi S. Shambharkar and Latesh G. Malik.
Supervision: Shubhangi S. Shambharkar and Latesh G. Malik.
Approval of the final text: Shubhangi S. Shambharkar and Latesh G. Malik.

VII. ACKNOWLEDGMENTS

The authors appreciate peers and reviewers' technical discussions, helpful suggestions, and critical evaluations that strengthened the theoretical framework and experimental validations. They also commend open-source researchers and dataset providers for offering accessible repositories for large-scale empirical testing and repeatability. Their open data and collaborative research enable sophisticated intrusion detection and adaptive cloud defense.

VIII. REFERENCES

[1] Aswini, J., Rekha, K. S., Rosaline, R. A. A., & Sivaneshkumar, A. (2025). Enhancing security in cloud computing systems using hybrid feature selection and ensemble-based machine learning for intrusion detection. *Evolving Systems*, 16(3). https://doi.org/10.1007/s12530-025-09725-6.

[2] Dugyala, R., Chithaluru, P., Ramchander, M., Kumar, S., Yadav, A., Yadav, N. S., Elminaam, D. S. A., & Alsekait, D. M. (2024). Secure cloud computing: leveraging GNN and leader K-means for intrusion detection optimization. *Scientific Reports*, 14(1). https://doi.org/10.1038/s41598-024-81442-7

[3] Rezaei, H., Taheri, R., Jordanov, I., & Shojafar, M. (2025). Federated RNN for Intrusion Detection System in IoT Environment Under Adversarial Attack. *Journal of Network and Systems Management*, 33(4). https://doi.org/10.1007/s10922-025-09963-8

[4] Chen, Y., Lin, F. Y., Tai, K., Hsiao, C., Wang, W., Tsai, M., & Sun, T. (2025). A near-optimal resource allocation strategy for minimizing the worse-case impact of malicious attacks on cloud networks. *Journal of Cloud Computing*, 14(1). https://doi.org/10.1186/s13677-025-00749-6

[5] Menezes, R. J., Jayarin, P. J., & Sekar, A. C. (2024). A bizarre synthesized cascaded optimized predictor (BizSCOP) model for enhancing security in cloud systems. *Journal of Cloud Computing*, 13(1). https://doi.org/10.1186/s13677-024-00657-1

[6] Singh, B., Indu, S., & Majumdar, S. (2025). Comparative Analysis of Intrusion Detection Models Using Quantum Machine Learning Techniques. *Circuits, Systems, and Signal Processing*, . https://doi.org/10.1007/s00034-025-03256-w

[7] Mohan, M., Tamizhazhagan, V., & Balaji, S. (2023). A Perspicacious Multi-level Defense System Against DDoS Attacks in Cloud Using Information Metric & Game Theoretical Approach. *Journal of Network and Systems Management*, 31(4). https://doi.org/10.1007/s10922-023-09776-7

[8] Sharma, A., & Singh, U. K. (2025). Cloud computing security assurance modelling through risk analysis using machine learning. *International Journal of System Assurance Engineering and Management*, 16(3), 1287-1300. https://doi.org/10.1007/s13198-025-02705-8

[9] Krishnaveni, S., Chen, T. M., Sathiyanarayanan, M., & Amutha, B. (2024). CyberDefender: an integrated intelligent defense framework for digital-twin-based industrial cyber-physical systems. *Cluster Computing*, 27(6), 7273-7306. https://doi.org/10.1007/s10586-024-04320-x

[10] Awad, Z., Zakaria, M., & Hassan, R. (2025). An enhanced ensemble defense framework for boosting adversarial robustness of intrusion detection systems. *Scientific Reports*, 15(1). https://doi.org/10.1038/s41598-025-94023-z

[11] Mallidi, S. K. R., & Ramisetty, R. R. (2025). A multi-level intrusion detection system for industrial IoT using bowerbird courtship Inspired feature selection and hybrid data balancing. *Discover Computing*, 28(1). https://doi.org/10.1007/s10791-025-09632-z.

[12] Alotaibi, N. S. (2024). Cloud guard: Optimizing intrusion detection for fortifying privacy defenses with an optimized self adaptive physics Informed neural network. *Peer-to-Peer Networking and Applications*, 17(6), 4378-4394. https://doi.org/10.1007/s12083-024-01799-4

[13] Neto, E. C. P., Iqbal, S., Buffett, S., Sultana, M., & Taylor, A. (2025). Deep learning for intrusion detection in emerging technologies: a comprehensive survey and new perspectives. *Artificial Intelligence Review*, 58(11). https://doi.org/10.1007/s10462-025-11346-z

[14] Farhat, S., Patel, A., Barros, A. L. B., & Bamhdi, A. M. (2025). MF2S-CID: A dynamic multi-model framework for scalable and interpretable intrusion detection. *International Journal of Information Security*, 24(4). https://doi.org/10.1007/s10207-025-01077-1

[15] Kandoussi, E. M., Houmairi, A., El Mir, I., & Bellafkih, M. (2024). Enhancing cloud security: harnessing bayesian game theory for a dynamic defense mechanism. *Cluster Computing*, 27(9), 12509-12526. https://doi.org/10.1007/s10586-024-04604-2.

[16] Ullah, I., Deng, X., Pei, X., Mushtaq, H., & Khan, Z. (2025). Securing internet of vehicles: a blockchain-based federated learning approach for enhanced intrusion detection. *Cluster Computing*, 28(4). https://doi.org/10.1007/s10586-024-04943-0

[17] Long, Z., Yan, H., Shen, G., Zhang, X., He, H., & Cheng, L. (2024). A Transformer-based network intrusion detection approach for cloud security. *Journal of Cloud Computing*, 13(1). https://doi.org/10.1186/s13677-023-00574-9

[18] Sudha, C., & Bolla, S. (2025). Intelligent intrusion forecasting framework for distributed environment using federated learning. *Cluster Computing*, 28(6). https://doi.org/10.1007/s10586-024-05012-2

[19] Gupta, R., & Alam, T. (2023). An efficient federated learning based intrusion detection system using LS2DNN with PBKA based lightweight privacy preservation in cloud server. *Multimedia Tools and Applications*, 83(15), 44685-44697. https://doi.org/10.1007/s11042-023-17401-7.

[20] Madhavan Nair Vasanthamma, H. V., & Thankaswamy, J. (2025). A hybrid stacked sparse autoencoder and LightGBM framework for high-performance intrusion detection in IoT networks. *Iran Journal of Computer Science*, . https://doi.org/10.1007/s42044-025-00309-w

[21] Thankappan, M., Rifâ-Pous, H., & Garrigues, C. (2024). A distributed and cooperative signature-based intrusion detection system framework for multi-channel man In-the-middle attacks against protected Wi-Fi networks. *International Journal of Information Security*, 23(6), 3527-3546. https://doi.org/10.1007/s10207-024-00899-9

[22] Louati, F., Ktata, F. B., & Amous, I. (2024). Big IDS: a decentralized multi agent reinforcement learning approach for distributed intrusion detection in big data networks. *Cluster Computing*, 27(5), 6823-6841. https://doi.org/10.1007/s10586-024-04306-9

[23] N, M., G, S. H. K., R, S., & NR, J. I. R. (2024). Federated transfer learning for intrusion detection system in industrial iot 4.0. *Multimedia Tools and Applications*, 83(19), 57913-57941. https://doi.org/10.1007/s11042-024-18379-6.

[24] doost P. A., Moghadam, S. S., Khezri, E., Basem, A., & Trik, M. (2025). A new intrusion detection method using ensemble classification and feature selection. *Scientific Reports*, 15(1). https://doi.org/10.1038/s41598-025-98604-w

[25] Karn, A. L., Ghanimi, H. M. A., Iyengar, V., Siddiqui, M. S., Alharbi, M. G., Alroobaea, R., Yousef, A., & Sengan, S. (2025). Applying the defense model to strengthen information security with artificial intelligence in computer networks of the financial services sector. *Scientific Reports*, 15(1). https://doi.org/10.1038/s41598-025-15034-4

[26] Sorour, S. E., Aljaafari, M., Shaker, A. M., & Amin, A. E. (2025). LSTM-JSO framework for privacy preserving adaptive intrusion detection in federated IoT networks. *Scientific Reports*, 15(1). https://doi.org/10.1038/s41598-025-95966-z

[27] Xu, S., Shi, Y., Shi, L., & Zhang, H. (2025). Efficient network defense policies via GNN-enhanced reinforcement learning. *The Journal of Supercomputing*, 81(8). https://doi.org/10.1007/s11227-025-07431-3.

[28] Malathy, N., Swvtha, A. T., Leela, T. B., & Raaman, A. (2025). Real-Time Intrusion Detection in IIoT Stream Data Using Window-Based Weighted Ensemble Techniques. *SN Computer Science*, 6(1). https://doi.org/10.1007/s42979-024-03597-4

[29] Zhang, Z., Das, A., Huang, G., & Baskiyar, S. (2025). CAT: A simple heterogeneous ensemble learning framework for network intrusion detection. *Peer-to-Peer Networking and Applications*, 18(4). https://doi.org/10.1007/s12083-025-02000-0

[30] Kantharaju, V., Suresh, H., Niranjanamurthy, M., Ansarullah, S. I., Amin, F., & Alabrah, A. (2024). Machine learning based intrusion detection framework for detecting security attacks in internet of things. *Scientific Reports*, 14(1). https://doi.org/10.1038/s41598-024-81535-3

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1030-1044. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3195

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

ENERGY CONSUMPTION PREDICTION IN 3D PRINTING USING ENHANCED HYBRID DEEP LEARNING MODELS

M. Jayakrishna*[1], M. Vijay[2], P. Manoj Kumar[3] and R. Suthan[4]

[1,2]Centurion University of Technology and Management, Paralakhemundi, Odisha, India.
[3]Department of Mechanical Engineering, Kangeyam Institute of Technology, Kangeyam - 638108, Tamil Nadu, India.
[4]Department of Mechanical Engineering, SRM Institute of Science and Technology, Ramapuram Campus, Chennai - 600089, Tamil Nadu, India

[1]https://orcid.org/0000-0003-1647-2750, [2]https://orcid.org/0000-0002-9226-8711
[3]https://orcid.org/0000-0003-3632-8369, [4]http://orcid.org/0000-0002-5022-6775

Email: *jayakrishnamakka555@gmail.com, pasupathimanojkumar@gmail.com, itsmesuthan@gmail.com

ARTICLE INFO

Article History
Received: December 26, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
3D printing,
energy consumption,
convolutional neural network,
3D printer data,
long-short term memory.

ABSTRACT

The 3D printing method is characterized as an energy-intensive technology with a major negative influence on the ecosystem and sustainability. Since then, both business and academics have turned their attention to the problem of 3D printing energy usage modelling, prediction, and optimization. However, prediction of energy consumption is a common issue in 3D printing. Thus, this research aims to forecast the energy consumption in 3D printing employing a novel Convolutional Neural Network with Long-Short Term Memory (CNN-LSTM). The developed model has five stages: data collection, data augmentation, pre-processing, feature extraction, and prediction. In data collection, Ender-3 pro 3D printer data were collected; whereas, in pre-processing, data normalization was performed by min-max normalization and missing data imputations was carried out using Mean or median. Moreover, feature extraction was done using Principal Component Analysis (PCA) method. Furthermore, in prediction phase, CNN-LSTM is employed for forecasting the energy usage. Moreover, the model's performance has been evaluated in regards to Mean Square Error (MSE), correlation, Root-MSE (RMSE), Normalized-MSE (NMSE), and Mean-Absolute Error (MAE). Furthermore, comparison has been performed to identify the effectiveness of the presented model over other methods.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The manufacturing sector remains crucial to the growth of the world economy via jobs, social engagement, and economic power. Nevertheless, manufacturing operations also have a negative influence on the surroundings, compensating for a significant amount (nearly 90%) of manufacturing power usage and CO2 emissions as 84%, which present significant problems for the globe in the present era, including biodiversity loss, resource scarcity, and global warming [1]. Moreover, for the years to come, it is anticipated that the world's energy consumption would rise. The primary source of the increased CO2 emissions that causes increase in sea levels and climate changes seems to be the fossil fuels combustion. After decades of decarbonization efforts, conventional fossil fuels remain account for 84.7% of global energy use.

Since its introduction in the earlier 1990s, additive manufacturing (AM) or 3D printing processes have advanced, grown more sophisticated, and changed the way things have been created and planned. Layer-by-layer manufacturing processes allow for the development of operationally stratified materials, the manufacture of components with complicated geometries, and the reduction of material waste. From becoming an instrument for testing to gradually being used for end-part manufacturing, it has gone a lengthy ways. To print actual functional components from numerous types and materials shapes, a number of fabrication methods have been created, including selective laser-melting (SLM), stereo-lithography (SLA), laser-engineered net-shaping (LENS), fused filament-fabrication (FFF), and selective laser-sintering (SLS). AM methods give component designers a good chance to generate a variety of parts because to their considerable design flexibility for generating end-use components.

Yet, the great level of design flexibility in manufacturing also makes the AM method more challenging in regards to intricate mechanical systems and distinctive design elements. A solitary AM manufacturing process may also create many parts simultaneously in certain multiple-part manufacturing techniques, including SLS and SLM. The anisotropic character of the components, porosity owing to inadequate fusing between nearby filaments, and deformation as a consequence of residual tension as an outcome of the AM procedure's quick cooling nature have been specific difficulties to be solved. It is essential to have a thorough grasp of the link among the printing settings and the resultant microstructures and AM component's material characteristics, along with the feedstock material's process ability (powder flow-ability and rheological properties).

The AM procedures, nevertheless, involve a number of process variables that may have an impact on the reliability of the finished parts, necessitating a thorough knowledge of a number of fields, including thermal-mechanical engagement, fluid dynamics, solid-liquid engagement, grain innovation, and material characteristics. In addition, AM demonstrates the importance of treating and refining material qualities of goods with reduced material usage and enables quick product fabrication from concepts [2]. As a result, AM is used in a extensive range of commercial and institutional settings, comprising automotive, aerospace, and dentistry equipment [3]. Thought has been given to the emerging environmental worries about durability, in specific about energy usage [4].

Consequently, it is critical to improve power consumption in AM equipment. Thus far, this element has received increasing scrutiny from manufacturers and scholars. AM had the ability to provide a higher product yield, which would increase energy consumption. Earlier on in the production chain, eco-design is required to assist manufacturers and engineers with decision-making, energy management, and process enhancement [5]. Yet, because the AM sub-systems provide so much information during the procedure, it is difficult to enhance energy management. Furthermore, material jetting, powder-bed fusion, vat photo-polymerization, Direct-Energy deposition (DED), sheet lamination, binder spraying, and material expansion are the seven classifications into which AM techniques are often divided [6].

Several technologies were created for every AM procedure to fulfil the growing need for printing capabilities in regards of design, materials, and effectiveness. Energy management concerns are becoming increasingly important for sustainable manufacturing as industrial AM technologies have been used by more sectors [7]. The complexity of an AM technology, nevertheless, makes it difficult to analyze and estimate its energy usage because it often consists of numerous subsystems with various sub-procedures. If the system's energy usage could be anticipated prior to the start of the procedure, methods for increasing energy efficiency may be developed. Hence, for improved energy regulation in the AM sector, it is essential to unearth energy-related facts and create precise prediction frameworks.

The industrial procedures energy efficiency is thought to be directly connected to both external variables and the process variables [5]. Moreover, computerized simulations have been a crucial component of AM development and optimization, since they save costly manufacturing procedure trial and error. Prior employing AM to create the desired part, finite element-predicated multi-physics simulation models (FEM) [8] have been built to mimic the AM process. FEM-based computations, nevertheless, take a lot of time and money to run. This gives rise to the drive to create a machine learning (ML)-based prediction tool, which can instantly provide the simulation outcome rather than doing pricey physics-based computations.

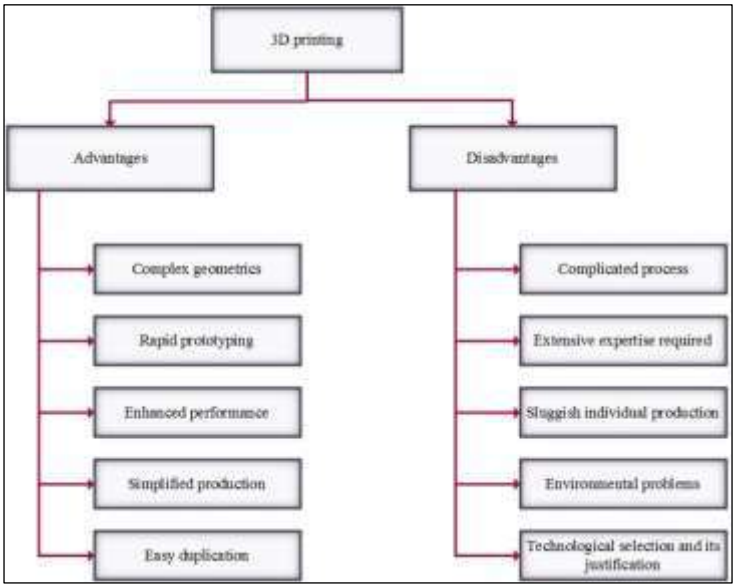


Figure 1: 3D printing Pros and Cons.
Source: Authors, (2026).

AM introduces a new manufacturing framework that produces tangible components by stacking materials in accordance with digitalized 3D models generated by computer-aided design (CAD) tools. The design possibilities of a design have been increased and the manufacturability has been enhanced by this free-form production method [9], [10]. Using a solitary 3D printer, it is possible to produce parts with specialized and intricate forms, accomplishing free-form production and eliminating the drawbacks of conventional manufacturing processes. In recent decades, additive manufacturing (AM) has demonstrated a significant promise for clean environmental and energy-saving operations because of mitigation in resources and material use and other equipment needs [11]. More information on AM procedures is now available thanks to the advent of data detection and gathering technology. The use of sophisticated data analytics methods to enhance energy management tactics has also increased.

The key important issue has been to retrieve and gain significant energy-relevant statistics and comprehension from this input from the present AM machines, which have been frequently integrated with various sensing equipment where there has been a tremendous quantity of data created during AM operations. Following that, the decision-making procedure is carried out using the data's information that has been retrieved [12]. When modelling with fixed structure, like neural networks, there is a trade-off among model reduction and effectiveness, therefore it's important to keep performance degradation to a minimum while preserving computing resources. The pros and cons of 3D printing are shown in Figure 1.Computational intelligence (CI) applications need to find solutions to the drawbacks of 3D printing.

This is particularly true when it comes to enhancing the fundamental printing processes by utilizing expertise from several fields such as 3D printer technologies, design and software engineering, and materials science. Overall else, the following have been necessary to find creative solutions to them: familiarity with material characteristics and interaction both during inside the object and printing; 3D printing innovation and material selection for a particular interface; design specifications for a particular printed artifacts; and parametric optimization [13], [14]. However, an AM technology often consists of numerous subsystems that each requires a separate sub-process, making it challenging to analyze the system's power usage. Thus, this research aimed to develop a novel hybrid deep learning (DL) for predicting the power consumption in 3D printing. The key contribution of the present work is described as follows:

- Initially, 3D printing model’s data like filament size, printing time, width, build speed, depth, and height are gathered.
- Hereafter, data augmentation process is performed to expand the training set by leveraging existing data to create updated duplicates of a database.
- Then the augmented data undergoes pre-processing in which the process done includes duplicate data removal, data normalization, and missing data imputation.
- Moreover, the features required for prediction are extracted using Principal Component Analysis (PCA).
- Finally, hybrid CNN-LSTM approach is presented, which predicts the energy consumption in 3D printing.

II. RELATED WORKS

AM represents a paradigm change in regards of operational adaptability and product customization, demonstrating considerable potential for industry-wide adoption. AM seems to be a rapidly growing innovation for rapid production. Because to the rising needs and uses of AM technologies in manufacturing, energy usage has recently drawn more interest in research and market. The energy usage of AM technologies, which are divided into multiple subsystems and are thought to be very complex, seems connected with a number of different aspects. These parameters come from numerous sources and frequently include properties of multiple kinds and dimensions, making it difficult to integrate them for analysis and modelling. By [15] presented a hybrid ML method to manage this multi-source information with distinct granularities and frameworks for forecasting energy usage in order to address this problem. This method integrates the density-based spatial clustering of applications with noise (DBSCAN) and extreme gradient boosting (XGBoost) method.

The article takes into consideration four distinct elements, namely working environment, design, method, and material. In order to decrease data complexity, the unorganized data has been grouped by DBSCAN and merged with custom attributes into the XGBoost algorithm for forecasting energy usage. To show the viability of the suggested approach, a case scenario that concentrated on the real-world SLS technology has been carried out. The findings showed that, like an ensemble learning method, XGBoost offers a workable method to forecast energy usage in a sophisticated AM system. Unfortunately, as a result of the lack of a feature extraction method, the projected and actual value exhibits a sizable discrepancy following prediction. In turn [16] looked at the design-relevant elements first in terms of energy modeling. Before manufacturing, process operators and part designers often decide on these aspects.

With the use of design-relevant data analysis, AM energy usage understanding that is concealed in the design-relevant characteristics is leveraged for predictive modeling. An innovative DL-driven particle swarm optimization (DLD-PSO) approach has been suggested to mitigate the energy consumption on the basis of the new modeling methodology. DL is used to improve the worldwide optimum of PSO and solve a number of concerns, including speeding up search times. Lastly, a case example is shown to test the suggested modelling technique, and the findings demonstrate its benefits, utilizing the design-relevant information gathered from a real-time AM technology in manufacturing. To help part manufacturers and process administrators rethink their plans and choices and lower the energy usage of the chosen AM technology under consideration, optimization was additionally conducted. However, the processing time is lengthy.

According to [17] used an ML-based technique to investigate various geometrical parameters at each printing phase and connect them to the mask image projection SLA's energy usage. By using layer-wise geometrical data from the design phase, the created frameworks would be capable of offering AM architects a practical tool for calculating power usage and raise knowledge of cleaner manufacturing in AM. Appropriate features have been chosen and/or retrieved from layer-wise geometric properties and utilized in the research to learn and evaluate ML models. The stacked autoencoders (SAE) architecture seems to have the highest testing effectiveness with 0.85% mean-RMSE, while the deep neural network does have 0.75% least mean RMSE (including both training and evaluation). Also, the DL approach performs better in testing than the neural network approach and regression approach, and it would be extremely useful in actual industrial situations.

However, additional environmental sustainability indicators, such as material use and operational emissions, have been not examined. AM has made it increasingly successful in the last ten years because the ability to produce functional components with complicated geometries quickly using methods like 3D printing and laser metal casting, which would be challenging to do with conventional machining. Computational methods were employed to mimic the performance of AM procedures prior to the trial run since the expense of producing a complex component for a costly metal like titanium becomes prohibitive. To forecast the AM procedures behavior, physics-informed data-driven ML algorithms are extremely helpful, but simulations have been time-consuming and operationally expensive for forecasting multi-scale multi-physics occurrences in AM. Such concepts enable real-world control techniques employing in-situ data along with multiscale simulation instruments.

Consequently, in order to create a data-driven method-based real-time control mechanism, [18] have created and built key elements of a methodological architecture. The dataset is created and time-dependent thermal equations have been solved using FE techniques. With the use of a computer program, the computer may be programmed to perform a variety of tasks, like calculating the time required to finish a task, calculating the number of steps required to finish a task, and more. The models estimate temperature distributions for AM procedures with an MAE of less than 1%. Although the focus of this study has been on temperature profile forecasting for an AM procedure, incremental forming cannot use the same concept. In the realm of manufacturing, AM seems to be a new manufacturing method that has the ability to manufacture complicated components and has great material usage, yet it additionally faces the issue of enormous energy usage, which has aroused worry.

The energy usage dispersion of AM apparatus has still been unexplored in the present study, which mostly concentrates on the AM energy usage procedure and its effects on the surroundings. By categorizing the apparatus into various energy units, [19] provided an analytical approach for resolving the AM equipment's energy usage distribution. The energy usage and dispersion of several AM equipment kinds, such as fused deposition modeling (FDM), SLA equipment, and SLM, have been studied in particular. The suggested energy usage quantification approach predicated on energy units has been then used to machine a typical structure in attempt to study the energy usage distribution features of the 3 distinct AM machines. The findings shown that the suggested technique can rapidly and accurately forecast how much energy AM machine will use. A procedure optimization technique taking energy usage and formation quality into account is given to achieve the best process variables for FDM predicated on the energy usage distribution approach.

This technique can enable the forecast of energy usage and the enhancement of AM's energy productivity. However, each energy unit's energy usage is not optimized. Although AM offers several benefits than subtractive manufacturing techniques, one of its main drawbacks was surface integrity. By [20] provide a data-driven forecasting modelling strategy to forecasting surface roughness in AM to enhance the additively built component's surface integrity. Heat and vibration information is gathered using a variety of sensors, includes infrared temperature sensors, accelerometers, and thermocouples. For the purpose of developing the surface roughness prediction framework, an ensemble learning approach has been presented. Features inside the frequency and time dimensions have been retrieved from sensor-based circumstance monitoring data. To increase the effectiveness of computing and the precision of predictions, a subgroup of these traits is chosen.

The circumstance tracking data gathered from a series of AM tests carried out on an FFF machine has been used to validate the prediction model. The suggested predictive modelling technique may accurately forecast the 3D printed element's surface roughness, according to experimental data. Nevertheless, this model cannot be modified to forecast the additively made component's surface roughness. The environment's constrained potential and the natural resource's depletion have made the sustainability challenges more urgent. Being the cornerstone of human civilization and society, production is crucial to sustainability. Layer-by-layer item fabrication is a novel method of production that offers a possible alternative to conventional subtractive manufacturing. The AM's sustainability performance hasn't been calculated or assessed, however, in a satisfactory manner. Most energy consumption research on AM in the research nowadays focuses on establishing links between process variables and power use.

Although these researches can help with integrated process engineering and ecological sustainability, they don't link product design to sustainability effectiveness, which limits their predictive power. As a result, they can't be utilized directly to help the creation and redesign of products. Moreover, connecting the needed power usage with the item shape can help with the creation of a life-cycle inventory dataset for AM operations. In attempt to assess and forecast the mask image representation SLA procedure's energy usage, [21] employed an innovative ML-based technique to extract geometry-related characteristics. The research's findings will be a crucial component of models for unit production processes and add to the AM's life-cycle inventory by addressing the disparity between process energy and usage product geometry. However, the computational time is high. The importance of attaining sustainable manufacturing continues rising as environmental consciousness and environmental legislation are being prioritized.

AM is a desirable method for creating environmentally friendly production. Unfortunately, a universal approach to predict the AM's energy usage is absent due to the variety of AM kinds and the varying working conditions of machines' parts. In order to anticipate energy usage, [22] created a novel model that took into account the energy of each part, the duration of each operation, and the functioning status of each element. A common AM method called FDM has been chosen to show how well the suggested model works. In comparison to the particular energy paradigm and the procedure-based power consumption approach, it has been discovered that the suggested framework had greater prediction accuracy. The suggested model may be quickly incorporated into the software to view the energy usage and printing time of each procedure in each element and, moreover, to serve as a benchmark for synchronizing the improvement of element quality and power usage.

However, the forecast method did not account for the time wanted for unfilled travel or the speed and nozzle's motion deceleration during the printing step. SLM's applications have been growing as a potential AM technique. Nevertheless, the SLM procedure consumes a lot of energy because of the intricate structures of SLM equipment and their slow processing speeds. For an effective assessment and decrease of SLM power use, energy prediction is essential. However, the SLM processes energy consumption is challenging to anticipate because to the variety of SLM equipment’s and their numerous working states. A unique technique to predict the SLM procedure's energy consumption was given by [23]. The power and temporal modeling of equipment sub-systems and sub-processes, respectively serve as the foundation of the suggested methodology. Prediction accuracy may be significantly increased by determining the operational conditions of SLM equipment components in each sub-process.

To show the viability of the suggested approach, two instances of aluminum components produced by an SLM process employing an SLM 280HL machine have been chosen. The suggested technique performs better in energy usage predictions than particular, stage-predicated, and subsystem-predicated energy benchmark methods, according to the findings. Nonetheless, the multi-laser machine's product geometries, laser jump duration, and scan automation remain unaffected. An essential, intriguing, yet difficult challenge in smart manufacturing was how to accurately anticipate energy usage in the 3D printing. In turn [24] presented a new deep network called 3DPECP-Net to handle this issue. In order to comply with the lower energy requirements of the programs to combat energy deficit and global warming, this work is crucial.

The multisource information is divided into three components, comprising pixel-, processing-, and motion-source, in accordance with the smart energy usage modelling for 3D printing. The suggested 3DPECP-Net, in contrast to existing alternatives, is capable of successfully fuse necessary statistics in 3D printing, carrying complementary information's full benefit of various data inputs for more precise prediction. Existing approaches, on the other hand, either only leverage the CAD model's geometric characteristics or simply incorporate the processing variables into the forecasting. To capture global spatial data and local attribute details, the multisource fusion frame (MSFF) layers numerous convolutional and transformer blocks. In addition, researcher suggested a continuous attention memory network (CAMN) to effectively retrieve and retain the consistency across layers, pushing the learning of a comprehensive depiction for 3D printing energy usage forecast. Using an internal dataset, several tests are run, and the outcomes showed that the suggested 3DPECP-Net overtakes state-of-the-art techniques in aspect of prediction effectiveness. However, the processing time of the method is high.

For the creation of components with a near-net form, laser-aided AM (LAAM) seems to be a crucial metallic 3D printing and reprocessing technique. To assess and improve the distribution of residual pressure and deformation that results, it is crucial to examine the thermal field that is created by various scanning procedures. Nevertheless, utilizing a current numerical framework to simulate the multi-bead layering method in order to assess and choose the best laser scanning tactics is highly computationally costly. In order to produce training information for a physics-based ML system, [25] make utilization of a newly formed and experimentally confirmed effective thermal field forecasting mathematical method for LAAM. To determine the relationship among laser scanning sequences and their accompanying thermal history dispersion, an integrated Recurrent Neural Networks (RNN) with Deep Neural Networks (DNN) (i.e., RNN-DNN) framework has been created. The presented RNN-DNN method can then estimate the thermal field for any shape employing a variety of scanning techniques. The RNN-DNN forecasts and the outcomes of the mathematical simulation demonstrated high concordance of greater than 95%. However, this method is not performed for multi-layered 3D-deposition procedure.

III. METHODOLOGY

The main purpose of the presented work has been to forecast the energy consumption in 3D printing. The presented method contains five major steps: data collection, data augmentation, data pre-processing, feature extraction and prediction. The PCA method is utilized for the feature extraction stage. Moreover, CNN-LSTM is employed for energy consumption prediction. The developed workflow is outlined in Figure 2.

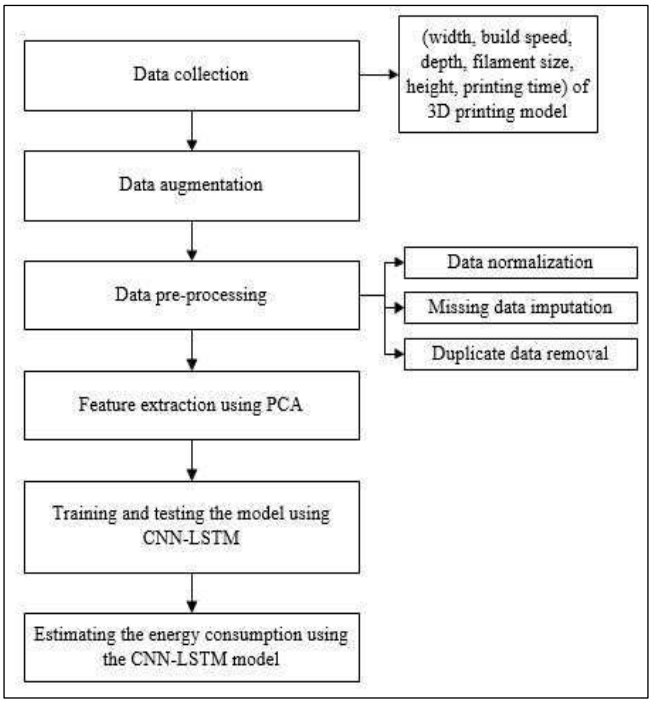


Figure 2: Workflow of presented research.
Source: Authors, (2026).

III.1 DATA COLLECTION

The intended 3D printing model has been employed to obtain the source data, in which the filament size, printing time, width, build speed, depth, and height are gathered. The data of Ender-3 pro 3D printer was used in this research. Design specifications for each tier, which have been frequently chosen at the outset of the whole process, are typically included in the information gathered from CAD drawings produced by designers.

III.2 DATA AUGMENTATION

The size of the corpus is necessarily an important attribute for a proper machine translation and the prediction of energy usage which is given as input. It helps in the development of an efficient data-driven procedure. For this purpose, the augmentation process is carried out in the study. In data analytics, procedures called "data augmentation" have been employed to expand the amount of information by introducing slightly changed versions of either existent data or brand-new artificial data that is derived from existent data.

While training a DL network, it functions as a classifier and aids in lowering overfitting. It has been strongly connected to data analytics oversampling. Moreover, by generating additional and divergent occurrences for training databases, data augmentation aids DL methods perform enhanced and yield better outcomes. A DL approach performs best and becomes more reliable if the database has been large and adequate. Moreover, Data Augmentation has been carried out while the original training set has been too small, to mitigate operating costs associated with identifying and cleansing the original database, to increase model reliability, and to avoid methods from overfitting.

III.3 DATA PRE-PROCESSING

The data have been pre-processed using a created framework and min-max normalization. Data stream mining techniques that may be used to prepare knowledge include knowledge normalization. In order to change the random number that correlates to the data, the values are raised until they fall within the predetermined points, like within 1.0. The Min-Max normalization approach has been employed to accomplish linear transformation beginning with the first stage. A rate of g of b is converted into g' within the range $[new_{\min(b)}, new_{\max(b)}]$ by min-max normalization. The component produces tuples with unfinished data by recommending among the numerous options, like majority, mean, constant, minimum, and variance, before performing the normalization technique to the supplemented data. The min-max normalization has been determined by applying Equation (1).

$$g' = {[g - \min(b)] \times [new_{\max(b)} - new_{\min(b)}]}/{[\max(b) - \min(b)]} + new_{\min(b)} \quad (1)$$

Here, $\min(b)$= attribute's minimum value, and $\max(b)$= attribute's maximum value. Equation (2) represents the formula mitigated for identifying the social control.

$$g' = {g - \min(b)}/{\max(b) - \min(b)} \quad (2)$$

A minimum and maximum normalization, or min-max normalization, upholds the relationship between the quantities of the original data.

Duplicate data removal: The divide between the training, verification, and testing sets may be destroyed by redundant data when comparable items are not all contained in a single set. This might lead to erroneous performance forecasts that let the framework down in terms of actual outcomes. As a result, the superfluous information is removed.

Missing data imputation: By replacing the missing data in the databases with another number, imputation appears to be a method for keeping the bulk of the information stored in a database. Since it's impractical to erase information from a database after each session, several techniques are used. By doing this, the size of the dataset will be drastically reduced, raising concerns about discrimination and leading to erroneous analysis. Mean or median interpolation substitutes all occurrences of missing data within a statistic with the parameter's mean or median.

III.4 PCA-BASED FEATURE EXTRACTION

To perform a full and rigorous analysis of the problems, a number of affecting factors should be evaluated in empirical studies. Yet, it gets significantly more challenging as the information assessment process grows more sophisticated. As the data provided by some inter-correlated components might overlap, the original information's relevant data might be expressed with fewer variables to decrease the level of complexity in the feature space. In accordance with this concept, Karl Pearson developed PCA in 1901, and Harold Hotelling independently came up with and updated it in the 1930s [26]. Based on the orthogonal transformation, PCA converts the possibly connected characteristic features into a distinctive group of completely uncorrelated data, referred as principal components. The number of primary components is either equivalent to or less than the original characteristic variables. Before to beginning the CNN-LSTM technique, PCA has been used to reduce the amount of data and produce a benchmark for swiftly making a prediction. The PCA's execution is defined as follows:

When a statistical framework $r \in D^q$ with q distinctive parameters and N occurrences is provided, it transforms into a completely other set of data $w \in D^h$:

$$w_{y,1} = (c_{11}\, r_{y,1} + c_{12}\, r_{y,2} + \cdots + c_{1q}\, r_{y,q}), (w_{y,2} = c_{21}\, r_{y,1} + c_{22}\, r_{y,2} + \cdots + c_{2q}\, r_{y,q}) \ldots\ldots (w_{y,r} = c_{r1}\, r_{y,1} + c_{r2}\, r_{y,2} + \cdots + c_{rq}\, r_{y,q}) \quad (3)$$

$$Here, \forall r_y = \left(r_{y,1}, r_{y,2}, \ldots\ldots, r_{y,q}\right)^t \in r, \qquad y = 1,2, \ldots\ldots, N \quad (4)$$

Every succeeding element does have the largest variance possible within its own set of limitations after the initial primary component $G_1 = w_{1,1}, w_{2,1}, \ldots\ldots, w_{N,1}$, which has the highest variance possible within the limitation that renders it transverse to the following components. The following are indicators of the PCA characteristics:

$$Variance(G_1) \geq Variance(G_2) \geq \cdots \geq Variance(G_h) \quad (5)$$

$$Covar(G_a, G_b) = 0, a \neq b, a, b = 1,2, \ldots\ldots, h \quad (6)$$

Where the $h \times q$ matrix k specifies the $y-th$ rows and a distinct coordinate structure and $k_y = k_{y1}, k_{y2}, \ldots\ldots, k_{yq}$ actually makes up element $y-th$ of the information covariance matrix M.

$$M = \frac{1}{N}\sum_{r=1}^{N} r_j r_j^t \tag{7}$$

In order to solve the characteristic formulation, PCA first computes the covariance matrix M:

$$Mk_y = \gamma_y k_y, \qquad y = 1,2, \ldots, h \tag{8}$$

Obtaining the related eigenvectors of the eigenvalues as well as the eigenvalues $\gamma_y (j = 1,2, \ldots, h)$. The eigenvector with the largest eigenvalue, G_1, which represents the most variation in the dataset r, makes up the initial component. This has been accompanied by the eigenvector with the second-largest eigenvalue, G_2, and so on. It might organize the eigenvalues in descending order and then choose the initial h main components to represent the real data in attempt to lessen the database size while keeping as much data as feasible.

III.5 CNN-LSTM BASED ENERGY CONSUMPTION PREDICTION

The CNN seems to be the deep neural network. Its objective is to identify the fundamental and inherent properties through the guided analysis of 2- or 3-dimensional data. These features have been suitable for predicting and identifying energy consumption in 3D printing. In a typical CNN architecture, an input layer is connected to several dense layers, fully-connected layers, pooling layers, an output layer, and a convolutional layer.

III.5.1 Convolutional Layer

This layer performs a convolution procedure using the raw input information and convolution kernels to produce new attribute variables. The approach was developed to extract attributes from a database; hence, the incoming data has to be structured into a matrix. The convolution kernel was already conceptualized as a tiny aperture that, in comparison to the input arrays, organizes coefficient values into a column. Convolutions are being performed on each layer while the display "slides" over the input arrays. A feature parameter called a convolved structure has been created by the filter's assigned dimension component and coefficient values. Applying different convolution operation to the input data will generate convoluted features, which have been frequently more helpful than the primary characteristics of the source data. Thus, the method works enhanced. The premise of CNN seems to be a convolutional layer, since the large number of calculations has been finished at this layer. It seems to be a data extraction stage that uses filters to extract local features, creates a feature map computed using convolution, leaves kernel function, also moves on to the pooling layer. Equation (9) contains convolutional layer expression.

$$P_m{}^{(a)} = \sigma(G_m{}^{(a)} + \sum_{n=1}^{s(a-1)} P_n{}^{(a-1)} * U_{m,n}{}^{(a)}) \tag{9}$$

Here, * (operator) represents convolution operation and $U_{m,n}{}^{(a)}$ represents filter linking the n^{th} feature-map in a-1 layer with the m^{th} feature map in a layer seems to be the function, which is utilized enhance nonlinearity, and activation matrix is denoted as σ.

III.5.2 Pooling Layer

Generally, pooling layer has been implemented after the convolutional layer. The information in the outcome of the convolution layer must be streamlined by the pooling layer. Moreover, the convolutional layer's feature map information from each feature map is used by the pooling layer to develop a condensed feature map. The most widely used strategies are mean and maximum pooling. On this stage, no learning has been taking place. In this level, size M×M filters were employed. Eqnuations (10) and (11) illustrates the average pooling and maximum pooling layers, correspondingly.

$$\bar{a} = \frac{1}{L}\Sigma_{(m,n)\in G} a_{m,n} \tag{10}$$

$$a_{max} = max_{(m,n)\in G}(a_{m,n}) \tag{11}$$

Where $a_{m,n}$ represents the quantity of each data in portion G and L denotes area's data count.

III.5.3 Dense Layer

The LSTM approach has been used in the dense layer. A subtype of recurrent neural networks with the ability to learn over time by using feedback links is the LSTM neural networks [27]. This technology generates STM and gets statistics from it utilizing cyclic linkages on their hidden layers and gather data from both time-series and sequences. Each LSTM unit consists of a memory cell as well as the three main gates for input, output, and forget.

This framework allows the LSTM to select the data that should be "forgotten" and the data that should be "remembered," resulting in a governed data stream and training for long-term interconnections. Equation (12) forecasts the LSTM unit's effective functioning. The developed approach's layers are given in Figure 3.

$$a_u = \sigma(P_u m_s + R_u n_{s-1} + t_u) \tag{12}$$

Here, weight matrices are R and P, m_s indicates input, σ represents the sigmoid function, and t denotes bias-term vector.

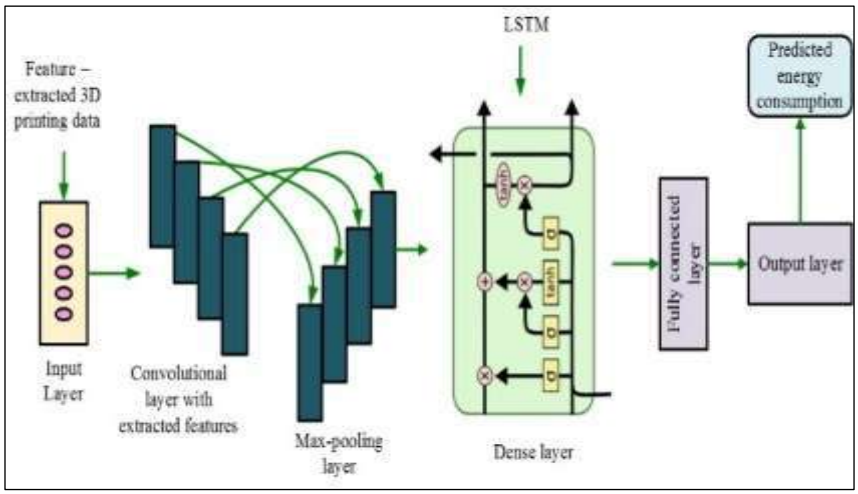


Figure 3: Developed CNN-LSTM's layers.
Source: Authors, (2026).

III.5.4 Output Layer

The output layer's cells, also referred to as the fully-connected layer, are totally dependent on all areas of the preceding layer of the memory. Behind this layer, information is turned into a 1-D array. The total number of completely linked layers in each design may vary. In this level, feed forward has been accomplished using Equation (13).

$$a_m^s = \sum_n u_f^{s-1} v_n^{s-1} \tag{13}$$

Here, number of layer is represented as s, v_n^s represents number in the developed output layer, number of neuron is represented as m and n, u_f^(s-1) indicates hidden layer's weight, v_n^(s-1) denotes input neuron and a_m^s denotes activation function's value in the output layer. The overall process of work is shown in algorithm 1.

Algorithm 1**:** CNN-LSTM

Input: Ender-3 3D printer data
Output: Energy consumption
Import input 3D printer data
Let X be the source data, which is considered for assessment
$X = \{X_1, X_2, X_3 \ldots.\}$
Data augmentation
Pre-processing //Min-max normalization, mean or median
Feature extraction //PCA
Initialize the CNN
Dense layer //LSTM
Prediction //CNN-LSTM
Predicting: 3D printer's energy usage
end if
end while
end

Source: Authors, (2026).

IV. RESULTS AND DISCUSSIONS

This section evaluates the suggested hybrid optimization with DL-based energy consumption performance and compares with conventional approaches. The outcomes of the simulation have been achieved for different models using different performance measures in order to evaluate the suggested DL model-based prediction's effectiveness.

Table 1: Parameter setup.

Parameters	Values
Molding method	FDM
Nozzle diameter	0.0004m
Model	Ender-3 pro
Printing size	0.22×0.22×0.25 m
Speed	50mm/s
Machine size	0.44×0.42×0.465 m
Printing precision	±0.1mm
Net weight	6980g
File transfer	SD-card offline or online
Filament	PLA
Printing temperature	210º

Source: Authors, (2026).

The suggested technique has been applied in MATLAB R2022b, and the experimental outcomes were obtained. The parameters employed for implementation is shown in Table 1. Moreover, the input models given for energy consumption prediction are shown in Figure 4 (a) and (b).

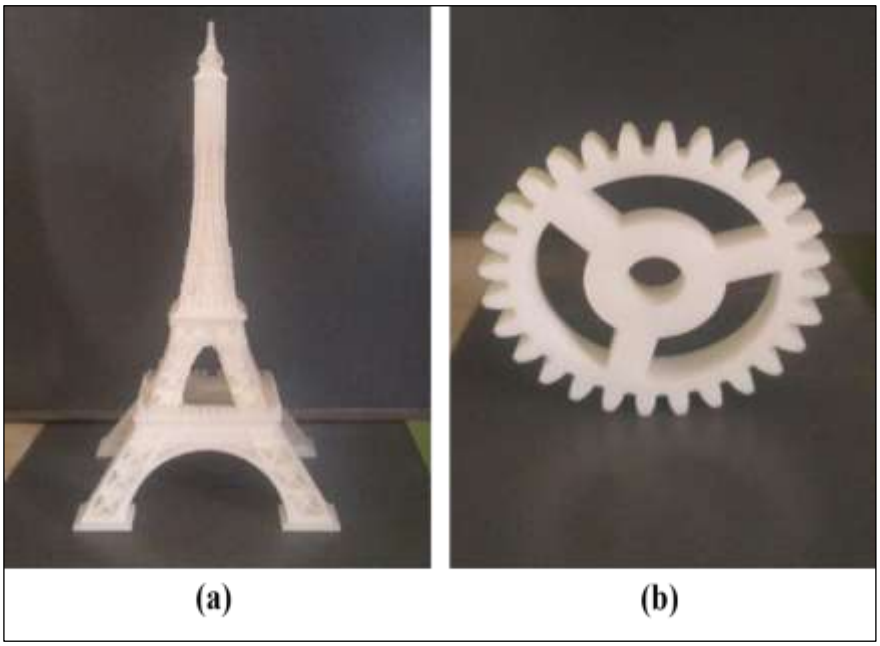


Figure 4: Input models (a) Eiffel tower (b) Gear.
Source: Authors, (2026).

IV.1 OTHER DL MODELS EMPLOYED FOR ASSESSMENT

IV.1.1 GRU

GRU is indeed a form of RNN, which was created by [28] as a simplified option to LSTM networks. GRU can analyze sequential data, including time-series data, text, and speech, just as LSTM. The fundamental principle of GRU would be to dynamically modify the network's hidden layer at each iteration interval via gating techniques. Information flow into and out of a system is managed by the gating procedures. The update gate and the reset gate are two of the GRU's 2 gating systems.

IV.1.2 CNN

In order to handle data with a grid pattern, like images, CNN seems to be a form of DL model that has been motivated by the structure of animal's visual cortex. From lower- to higher-level structures, CNN has been designed to continually and aggressively gather spatial categories of attributes. A typical CNN is made up of three different types of layers: fully-connected, convolution, and pooling layers. Convolution- and pooling-layers in order each and two provides features extraction, whereas a fully-connected layer in order multiple transforms the obtained characteristics into the consequence, such forecasting. A convolution layer appears essential in CNN that is composed of a series of mathematical operations, comprising convolution, a particular form of linear component [29].

IV.1.3 LSTM

LSTMs are employed in the area of DL. A number of RNNs, particularly in issues involving sequence forecasting, are susceptible of acquiring long-term connections. In addition to solitary data sources like images, LSTM also has feedback links that let it to analyze the whole data stream. This is useful for speech recognition and machine translation, among other things. The LSTM is a special type of RNN that excels at solving a variety of problems.

IV.1.4 Bi-LSTM

Two LSTMs—one receiving the input towards the forward manner while the second processes it in a backward direction—combine to form the Bi-LSTM, a sequential processing paradigm. With the assistance of Bi-LSTMs, the system has entrance to more statistics, which assistances the method's setting. One extra LSTM layer is added by Bi-LSTM, which changes the data flow's orientation. That simply implies that in the extra LSTM block, the incoming data flows backwards. The outcomes from the 2-LSTM layers have been then combined employing a numerous methods, comprising mean, sum, concatenation or multiplication. Many benefits come with this kind of design for solving real-world issues.

IV.2 DL MODEL'S PERFORMANCE EVALUATION

In order to assess the energy consumption forecast, the various DL techniques' performance metrics have been looked at. The MSE, MSE, MAE, NMSE, correlation, and RMSE have all been used in this study to assess how well the DL model has worked. To compute prediction errors and assess prediction models, several metrics are used. The MAE, MSE, and RMSE are well-known scale-dependent metrics based on real and squared values.

IV.2.1 MSE

The mean square error (MSE) appears to be measured in statistics. It is a risk function that corresponds to predicted values. It could have never been null since it is arbitrary. It has consistently been employed to evaluate estimate performance in a positive way. MSE appears to be a secondary instant that takes estimate variance and error causes into account. The MSE seems to be an unbiased predictor with estimate directionality. It has been calculated using the mean squared variances between the actual and anticipated values.

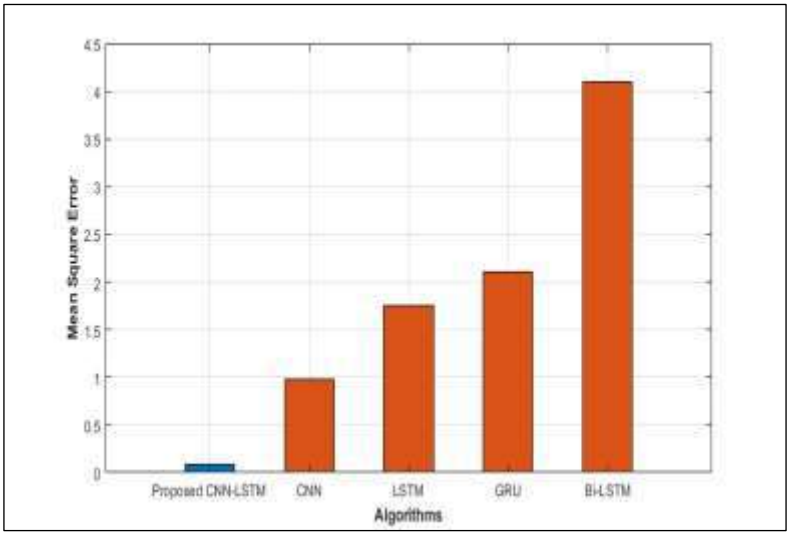


Figure 5: Obtained MSE.
Source: Authors, (2026).

The above Figure 5 indicates the graphical chart of MSE. The result indicated that the presented CNN-LSTM model has attained less MSE of 0.080762 when compared to other method like CNN, Bi-LSTM, GRU, and LSTM. Moreover, the attained MSE of CNN, Bi-LSTM, GRU, and LSTM are 0.979475, 4.10168, 2.099311, and 1.754408, correspondingly. It is evident that the presented method has less prediction error.

IV.2.2 RMSE

The MSE appears to be a statistic that is frequently used in estimates. The RMSE was utilized instead of the MSE by taking the square root. The RMSE calculates how far an estimate deviates from reality. Equation (14) calculates RMSE.

$$\mathrm{RMSE} = \sqrt{{}^{1}/_{\mathrm{N}} \times \sum_{u=1}^{\mathrm{N}} (\mathrm{A_u} - \mathrm{D_u})^2} \tag{14}$$

Where, $|\mathrm{A_u} - \mathrm{D_u}|$ is the absolute error, and N is the number of errors.

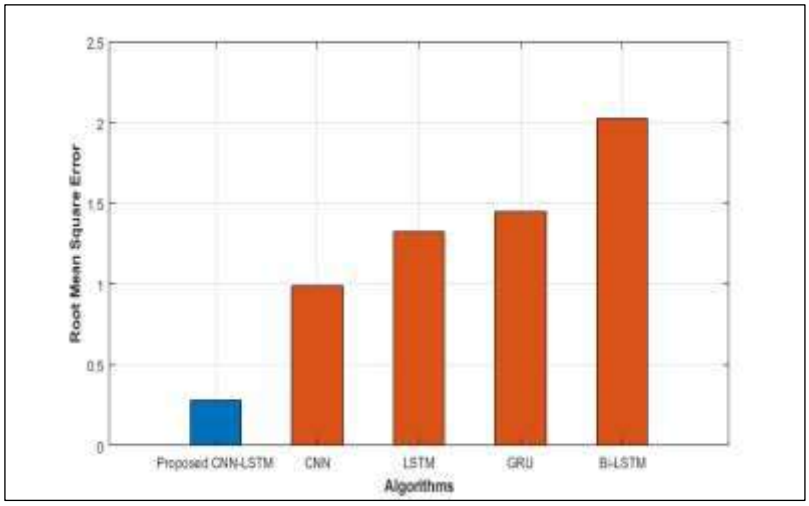


Figure 6: Obtained RMSE.
Source: Authors, (2026).

The above Figure 6 presents the graphical chart of RMSE. The result indicated that the presented CNN-LSTM model has attained less RMSE of 0.284187 when compared to other method like CNN, LSTM, GRU, and Bi-LSTM. Moreover, the attained MSE of CNN, LSTM, GRU, and Bi-LSTM are 0.989, 1.324, 1.448, and 2.025, correspondingly.

IV.2.3 MAE

By considering the average amount of error over a set of predictions while disregarding the direction of the mistake, the MAE assigns less weight to exceptional predictions. Equation (15) calculates MAE.

$$\mathrm{MAE} = {\sum_{u=1}^{N} |A_u - D_u|}/{N} \tag{15}$$

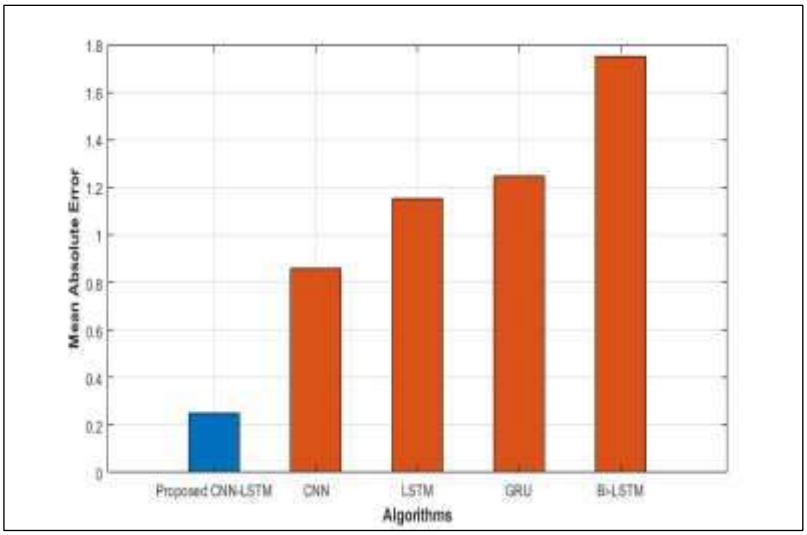


Figure 7: Obtained MAE.
Source: Authors, (2026).

The above Figure 7 present the graphical chart of MAE in which the presented approach has been contrasted with other different method like CNN, Bi-LSTM, GRU and LSTM. From the graph it can be stated that the presented method (CNN-LSTM) has low MAE of 0.251005 and Bi-LSTM has high MAE range. This makes the proposed method more effective.

IV.2.4 NMSE

When the training procedure has been executed for every instance in the training database, one epoch happens. Finding the weights, which reduce a particular total error measure, like the sum of MSE or squared errors, seems to be the ultimate objective. The commonly utilized NMSE is utilized in this study to assess the precision of estimates of energy consumption and is described as follows:

$$\mathrm{NMSE} = {1}/{\sigma^2 N} \times \sum_{u=1}^{N} (A_u - D_u)^2 \tag{16}$$

Where, σ^2 represents the time-series variance.

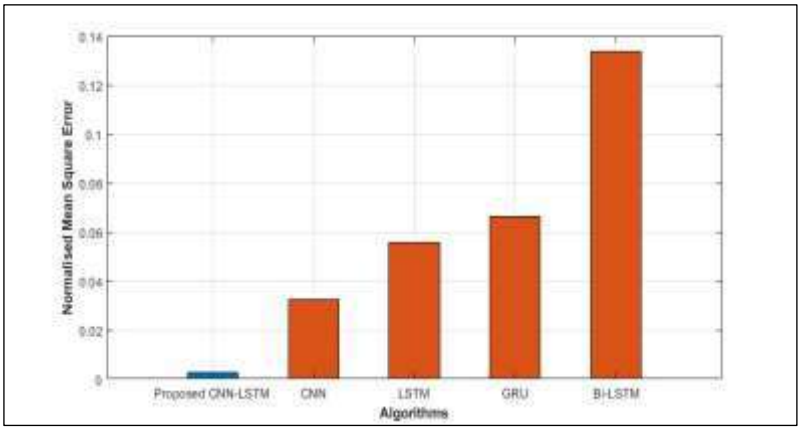


Figure 8: Obtained NMSE.
Source: Authors, (2026).

The above Figure 8 present the graphical chart of NMSE in which the presented model has been contrasted with other different method like CNN, Bi-LSTM, GRU and LSTM. From the graph it can be stated that the presented method (CNN-LSTM) has low NMSE of 0.00272 and Bi-LSTM has high NMSE range. Since the NMSE range is low the performance could be more effective.

IV.2.5 Correlation

The relationship between parameters or attributes in a dataset is quantified by correlation coefficients. There isn't any inclination for the levels of the parameters to rise or fall together when there has been no connection between them. Even if two variables have been uncorrelated, they could still not be autonomous because of a nonlinear connection.

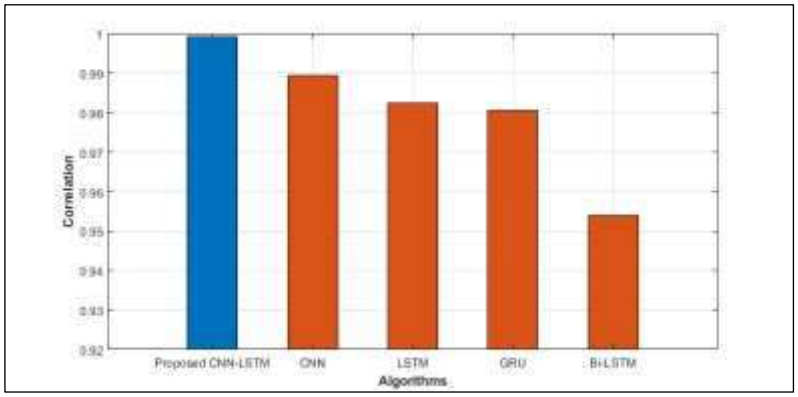


Figure 9: Obtained correlation.
Source: Authors, (2026).

The above Figure 9 present the graphical chart of correlation in which the presented approach was contrasted with other different model like CNN, Bi-LSTM, LSTM and GRU. From the graph it can be stated that the presented method (CNN-LSTM) has high correlation range of 0.999212 and Bi-LSTM has low correlation range about 0.95394. The method with higher correlation is suitable for accurate prediction. Thus, the developed model has higher prediction reliability.

IV.3 PERFORMANCE EVALUATION OF PRESENTED CNN-LSTM

For the prediction of energy consumption, accurate prediction is crucial. Since the results of the obtained prediction have not been very excellent, it would appear that developing a more trustworthy method is essential.

Table 2: Performance of the proposed method CNN-LSTM.

Evaluation metrics	Proposed CNN-LSTM
MAE	0.251005
RMSE	0.284187
MSE	0.080762
NMSE	0.00272
Correlation	0.999212

Source: Authors, (2026).

When evaluating the performance of the proposed technique, the MAE, RMSE, MSE, NMSE, and Correlation were all taken into account. Table 2 and Figure 10 displays the outcomes of these variables used in the study, which made use of 80% training data and 20% testing data.

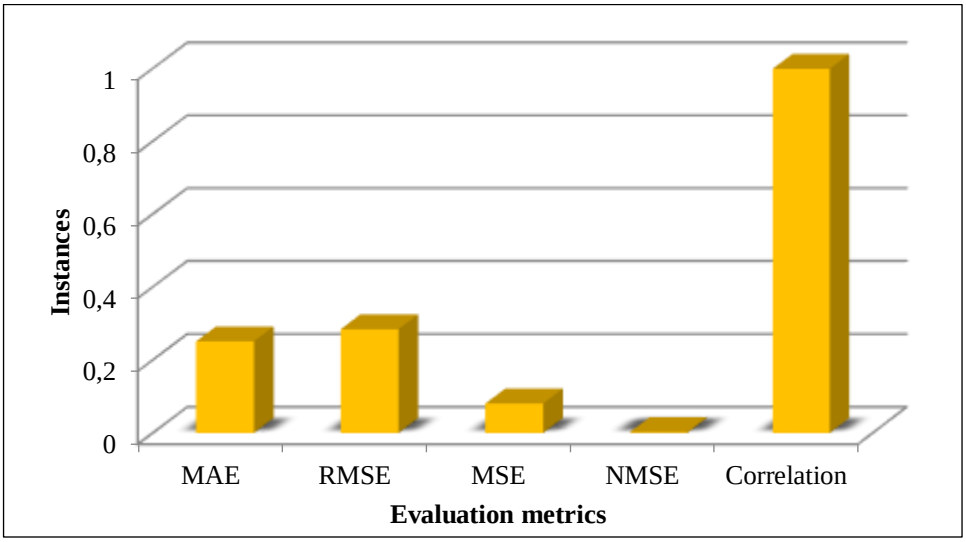


Figure 10: CNN-LSTM model performance.
Source: Authors, (2026).

The MAE, RMSE, MSE, NMSE, and Correlation of the provided CNN-LSTM model are all 0.251005, 0.284187, 0.080762, 0.00272, and 0.999212, respectively. According to the findings, the provided model has achieved less error in the prediction of energy consumption.

IV.4 COMPARISON

To determine the viability of the offered methods, proposed CNN-LSTM models were evaluated and compared to other methods already in use. The RMSE comparisons of current strategy with various approaches are shown in Table 3, and Figure 11. The presented DL model was compared with existing models like Support Vector-Regression (SVR), Gradient-Boosting Regression Tree (GBRT), CNN, and XGBoost [15].

Table 3: Comparison of RMSE.

Method	RMSE
SVR	154.79
GBRT	163.23
CNN	231.96
XGBoost	130.78
Proposed CNN-LSTM	0.284

Source: Authors, (2026).

The result indicated that the presented CNN-LSTM has attained less RMSE of 0.284 when compared to other state-of-art methods. The SVR method has obtained RMSE as 154.79, GBRT model has attained RMSE as 163.23, CNN approach has attained RMSE as 231.96, and XGBoost has attained RMSE as 130.78. From the comparison outcome, it is evident that the presented model's RMSE has enhanced up to 90% than the other existing methods.

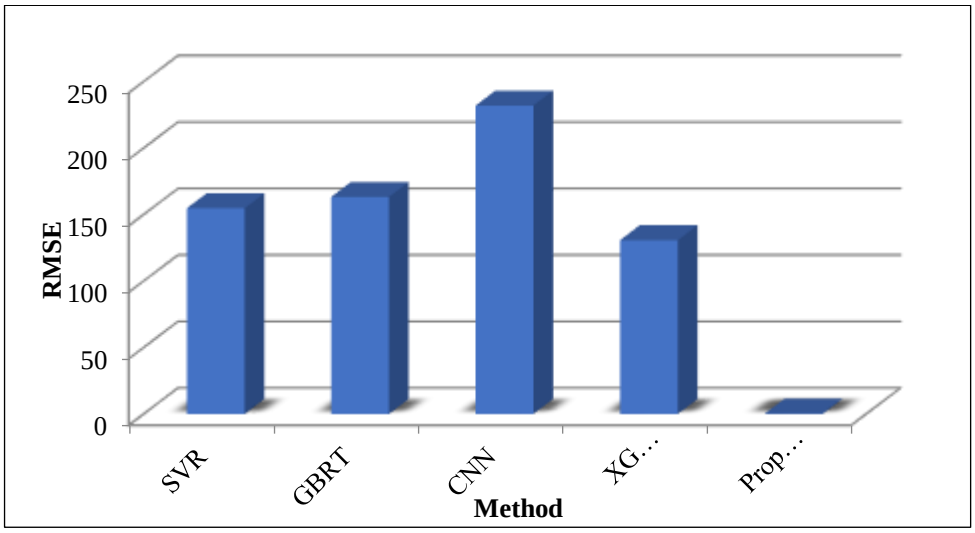


Figure 11: RMSE comparison.
Source: Authors, (2026).

V. CONCLUSIONS

AM seems to be a comprehensive manufacturing method that integrates a number of technologies and is presently used in a wide range of sectors. However, in AM techniques, energy use has grown to be a big problem. As a result, a hybrid DL (CNN-LSTM) strategy is suggested. A study of relevant studies highlighting the importance of CNN and LSTM in the industrial arena served as the inspiration for the suggested strategy. Data related to printing are used as source in a suggested DL-based energy consumption forecasting technique. Moreover, to enlarge the database, data augmentation process has been performed. In addition, pre-processing was performed to remove the duplicate data, missing data imputation, and data normalization. Hereafter, feature extraction was done by PCA that extracts the required features from pre-processed data for prediction. A novel hybrid CNN-LSTM model has been finally implemented for predicting the energy consumption. The result indicated that the presented model has attained less prediction errors while contrasted to other existing and state-of-art methods. Thus, the presented CNN-LSTM model has higher prediction reliability. However, this method employs smaller parameters for predicting 3D printing's energy consumption. Thus, in future variety of data kinds will be used for predicting energy consumption using optimization method with different hybrid DL framework.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: M. Jayakrishna and M. Vijay.
Methodology: M. Jayakrishna and P. Manoj Kumar.
Investigation: M. Jayakrishna and M. Vijay.
Discussion of results: M. Jayakrishna and P. Manoj Kumar.
Writing – Original Draft: M. Jayakrishna and M. Vijay.
Writing – Review and Editing: P. Manoj Kumar and R. Suthan
Resources: M. Jayakrishna and M. Vijay.
Supervision: M. Vijay
Approval of the final text: M. Jayakrishna, M. Vijay, P. Manoj Kumar and R. Suthan.

VII. REFERENCES

[1] J. R. Duflou et al., "Towards energy and resource efficient manufacturing: A processes and systems approach," CIRP Ann., vol. 61, no. 2, pp. 587–609, 2012, doi: 10.1016/j.cirp.2012.05.002.

[2] L. Lidong and C. A. Alexander, "Additive Manufacturing and Big Data," Int. J. Math. Eng. Manag. Sci., vol. 1, no. 3, pp. 107–121, Dec. 2016, doi: 10.33889/IJMEMS.2016.1.3-012.

[3] K. Kellens, M. Baumers, T. G. Gutowski, W. Flanagan, R. Lifset, and J. R. Duflou, "Environmental Dimensions of Additive Manufacturing: Mapping Application Domains and Their Environmental Implications: Environmental Dimensions of Additive Manufacturing," J. Ind. Ecol., vol. 21, no. S1, pp. S49–S68, Nov. 2017, doi: 10.1111/jiec.12629.

[4] Y. Yang, L. Li, Y. Pan, and Z. Sun, "Energy Consumption Modeling of Stereolithography-Based Additive Manufacturing Toward Environmental Sustainability," J. Ind. Ecol., vol. 21, no. S1, Nov. 2017, doi: 10.1111/jiec.12589.

[5] D. Freitas, H. A. Almeida, H. Bártolo, and P. J. Bártolo, "Sustainability in extrusion-based additive manufacturing technologies," Prog. Addit. Manuf., vol. 1, no. 1–2, pp. 65–78, Jun. 2016, doi: 10.1007/s40964-016-0007-6.

[6] I. Gibson, D. Rosen, B. Stucker, and M. Khorasani, Additive Manufacturing Technologies. Cham: Springer International Publishing, 2021. doi: 10.1007/978-3-030-56127-7.

[7] R. Huang et al., "Energy and emissions saving potential of additive manufacturing: the case of lightweight aircraft components," J. Clean. Prod., vol. 135, pp. 1559–1570, Nov. 2016, doi: 10.1016/j.jclepro.2015.04.109.

[8] W. Yan et al., "Data-driven multi-scale multi-physics models to derive process–structure–property relationships for additive manufacturing," Comput. Mech., vol. 61, no. 5, pp. 521–541, May 2018, doi: 10.1007/s00466-018-1539-z.

[9] T. Peng, K. Kellens, R. Tang, C. Chen, and G. Chen, "Sustainability of additive manufacturing: An overview on its energy demand and environmental impact," Addit. Manuf., vol. 21, pp. 694–704, May 2018, doi: 10.1016/j.addma.2018.04.022.

[10] Ch. Achillas, D. Aidonis, E. Iakovou, M. Thymianidis, and D. Tzetzis, "A methodological framework for the inclusion of modern additive manufacturing into the production portfolio of a focused factory," J. Manuf. Syst., vol. 37, pp. 328–339, Oct. 2015, doi: 10.1016/j.jmsy.2014.07.014.

[11] A. Majeed et al., "A big data-driven framework for sustainable and smart additive manufacturing," Robot. Comput.-Integr. Manuf., vol. 67, p. 102026, Feb. 2021, doi: 10.1016/j.rcim.2020.102026.

[12] K. Zhou and S. Yang, "Understanding household energy consumption behavior: The contribution of energy big data analytics," Renew. Sustain. Energy Rev., vol. 56, pp. 810–819, Apr. 2016, doi: 10.1016/j.rser.2015.12.001.

[13] I. Rojek, D. Miko\lajewski, P. Kotlarz, M. Macko, and J. Kopowski, "Intelligent system supporting technological process planning for machining and 3D printing," Bull. Pol. Acad. Sci. Tech. Sci., vol. 69, no. 2, 2021, doi: 10.24425/bpasts.2021.136722.

[14] I. Rojek, D. Mikołajewski, E. Dostatni, and M. Macko, "AI-Optimized Technological Aspects of the Material Used in 3D Printing Processes for Selected Medical Applications," Materials, vol. 13, no. 23, p. 5437, Nov. 2020, doi: 10.3390/ma13235437.

[15] Y. Li, F. Hu, J. Qin, M. Ryan, R. Wang, and Y. Liu, "A Hybrid Machine Learning Approach for Energy Consumption Prediction in Additive Manufacturing," in Pattern Recognition. ICPR International Workshops and Challenges, vol. 12664, A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, and R. Vezzani, Eds. Cham: Springer International Publishing, 2021, pp. 622–636. doi: 10.1007/978-3-030-68799-1_45.

[16] J. Qin, Y. Liu, R. Grosvenor, F. Lacan, and Z. Jiang, “Deep learning-driven particle swarm optimisation for additive manufacturing energy optimisation,” J. Clean. Prod., vol. 245, p. 118702, Feb. 2020, doi: 10.1016/j.jclepro.2019.118702.

[17] Y. Yang, M. He, and L. Li, “Power consumption estimation for mask image projection stereolithography additive manufacturing using machine learning based approach,” J. Clean. Prod., vol. 251, p. 119710, Apr. 2020, doi: 10.1016/j.jclepro.2019.119710.

[18] A. Paul et al., “A Real-Time Iterative Machine Learning Approach for Temperature Profile Prediction in Additive Manufacturing Processes,” in 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Washington, DC, USA, Oct. 2019, pp. 541–550. doi: 10.1109/DSAA.2019.00069.

[19] Z. Ma et al., “Energy consumption distribution and optimization of additive manufacturing,” Int. J. Adv. Manuf. Technol., vol. 116, no. 11–12, pp. 3377–3390, Oct. 2021, doi: 10.1007/s00170-021-07653-8.

[20] Z. Li, Z. Zhang, J. Shi, and D. Wu, “Prediction of surface roughness in extrusion-based additive manufacturing with machine learning,” Robot. Comput.-Integr. Manuf., vol. 57, pp. 488–495, Jun. 2019, doi: 10.1016/j.rcim.2019.01.004.

[21] Y. Yang, M. He, and L. Li, “A new machine learning based geometry feature extraction approach for energy consumption estimation in mask image projection stereolithography,” Procedia CIRP, vol. 80, pp. 741–745, 2019, doi: 10.1016/j.procir.2019.01.012.

[22] Z. Yan et al., “A New Method of Predicting the Energy Consumption of Additive Manufacturing considering the Component Working State,” Sustainability, vol. 14, no. 7, p. 3757, Mar. 2022, doi: 10.3390/su14073757.

[23] J. Lv, T. Peng, Y. Zhang, and Y. Wang, “A novel method to forecast energy consumption of selective laser melting processes,” Int. J. Prod. Res., vol. 59, no. 8, pp. 2375–2391, Apr. 2021, doi: 10.1080/00207543.2020.1733126.

[24] K. Wang, L. Yu, J. Xu, S. Zhang, and J. Qin, “Energy consumption intelligent modeling and prediction for additive manufacturing via multisource fusion and inter-layer consistency,” Comput. Ind. Eng., vol. 173, p. 108720, Nov. 2022, doi: 10.1016/j.cie.2022.108720.

[25] K. Ren, Y. Chew, Y. F. Zhang, J. Y. H. Fuh, and G. J. Bi, “Thermal field prediction for laser scanning paths in laser aided additive manufacturing by physics-based machine learning,” Comput. Methods Appl. Mech. Eng., vol. 362, p. 112734, Apr. 2020, doi: 10.1016/j.cma.2019.112734.

[26] H. Hotelling, “Analysis of a complex of statistical variables into principal components.,” J. Educ. Psychol., vol. 24, no. 6, p. 417, 1933.

[27] I. E. Livieris, E. Pintelas, and P. Pintelas, “A CNN–LSTM model for gold price time-series forecasting,” Neural Comput. Appl., vol. 32, no. 23, pp. 17351–17360, Dec. 2020, doi: 10.1007/s00521-020-04867-x.

[28] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, “On the Properties of Neural Machine Translation: Encoder-Decoder Approaches,” 2014, doi: 10.48550/ARXIV.1409.1259.

[29] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” Insights Imaging, vol. 9, no. 4, pp. 611–629, Aug. 2018, doi: 10.1007/s13244-018-0639-9.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1045-1053. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3209

RESEARCH ARTICLE OPEN ACCESS

FEDERATED LEARNING FOR ENHANCING RELIABILITY AND SECURITY IN MEDICAL IMAGE ANALYSIS AGAINST ADVERSARIAL THREATS

Meenu Mohil*[1], Akshita Mohil[2] and Abhiraj Singh Mohil[3]

[1]Department of Physics, Acharya Narendra Dev College, University of Delhi, Delhi 110019, India
[2, 3]Department of Computer Science, Netaji Subhas University of Technology, Delhi 110075, India

[1]http://orcid.org/0009-0005-7978-7217, [2]http://orcid.org/0009-0000-7882-9511, [3]http://orcid.org/0009-0000-0205-3550

Email: *meenumohil@andc.du.ac.in, mohil.akshita@gmail.com, mohil.abhiraj@gmail.com

ARTICLE INFO

Article History
Received: December 30, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Federated Learning,
Adversarial Robustness,
Privacy-Preserving AI,
Medical Image Analysis

ABSTRACT

Federated Learning (FL) has emerged as an attractive concept of facilitating privacy-saving artificial intelligence in the field of medicine. This paper suggests a secure and robust Federated Learning framework that enhances the robustness and privacy of medical image analysis, as well as safeguard the data confidentiality of decentralised healthcare organizations. The framework incorporates superior security protocols to protect the federated learning systems to model poisoning and adversarial interruptions. It enhances the resilience of the global model to inconsistent or malicious contributions of data by using adaptive aggregation, anomaly detection, and dynamic reputation-based client evaluation. It is more accurate, stable, and robust to use in a heterogeneous and non-IID environment, which guarantees the reliable cooperation of decentralized medical applications. The proposed structure comprises of a well-organized pipeline that involves data preprocessing, local training, and secure model aggregation. A convolutional neural network (CNN) is trained on histopathological images by each client and does not violate data privacy on an institutional level. The improved efficiency and reliability as well as the resilience of the model to disruptive hostile training of histopathological images are experimentally demonstrated on a curated dataset of histopathological images. It can be proved that the proposed system is well-diagnostic with enhanced compliance with privacy and offers a flexible and dependable model of healthcare partnership on a large scale. This paper highlights how federated learning can advance improved clinical outcomes and enhance the quality of patient care by demonstrating the result of applying federated learning to multi-institutional image analysis and showing improvements in model security and performance.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

I.1 THE ROLE OF DATA PRIVACY IN MODERN HEALTHCARE

In the era of massive data expansion and technological progress, the protection of data privacy is an important issue, particularly in the medical field, where patient data confidentiality is at stake. Federated learning continues to become hailed as a groundbreaking approach to machine learning whereby the cooperation of organisations can create the models without necessarily breaching the privacy of sensitive data [1]. Rather than centralizing patient data, each of the participating institutions retains the power over information and considerably minimizes the potential threats of unauthorized access to data or information breach and meet ethical and legal demands of confidentiality [2]. The question is still how to use the technological innovations as artificial intelligence (AI) becomes an inseparable component of clinical diagnostics and decision-making and how these factors would not threaten the confidentiality of patients and disregard the international standards of patient privacy [3].

I.2 FEDERATED LEARNING: A DECENTRALIZED APPROACH TO SECURE COLLABORATION

Federated Learning (FL) is a decentralized method which enables a group of institutions to share the operations of training machine learning models without sharing raw data across one organization to another. Under this setup, each client is learning a model on its own confidential data and only send back the new parameters in a central server where they can be aggregated together using a series of aggregation algorithm like the Federated Averaging (FedAvg) algorithm, as shown in Figure 1. The process allows the development of globally applicable models that will ensure the generalization of data among different healthcare settings, without the leakage of information [4]. FL has achieved success in diagnostic imaging, classification of the disease and customized prognostication. Further research on this area has been made by the introduction of attention-guided networks, adaptive aggregation, and reputation-based weight optimization to improve model robustness and trustworthiness in the application of AI in medicine [5], [6].

I.3 APPLICATIONS AND FUTURE PROSPECTS IN CANCER HISTOPATHOLOGY

Federated learning can be applied to cancer histopathology, which has complicated tissue structures and diagnostic issues, to a significant effect because, by using multi-institutional datasets, privacy of patient data can be preserved. Based on decentralized training models, histopathological data in various labs are combined to train convolutional and transformer-based neural networks, which improve the accuracy and reliability of the classification in identifying cancers. New FL architectures have reached near the same level of parity as centralized learning, and do not infringe the data protection legislation [7]. Besides privacy protection, FL reduces bias in the datasets and establishes inclusive healthcare innovation by enabling resource-constrained institutions to contribute. In addition to privacy protection, FL minimises bias in the datasets and creates inclusive healthcare innovation by allowing resource-constrained institutions to contribute. Federated learning is increasingly positioned as a core technology for ethical AI implementation in precision oncology and digital pathology [6], [8].

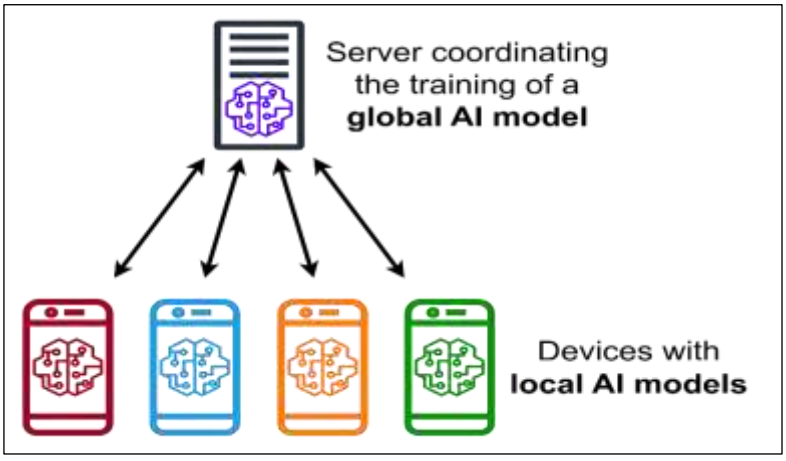


Figure 1: Outline of Federated Learning.
Source: Authors, (2026).

II LITERATURE SURVEY

This section reviews current developments in Federated Learning (FL) within the healthcare domain with particular attention to its application in medical image analysis and its capacity to enhance model performance and reliability. It also examines the security vulnerabilities inherent in FL and outlines the defense strategies proposed to counter adversarial threats in distributed learning environments.

II.1 FEDERATED LEARNING: CONCEPTS AND FRAMEWORKS

Federated Learning (FL) has become one of the most innovative models in the sphere of decentralized model training and the possibility to empower many medical institutions to develop strong AI systems without sharing sensitive patient data. In contrast to the idea of the conventional teaching methods of centralized learning when all the data is located in a single location, data sovereign is the feature of federated learning in which the raw data is stored on the local devices and only the model parameters are sent there to facilitate aggregation [1], [9]. This form of decentralization allows FL to be particularly useful to healthcare applications that require highly regulated privacy of data like the Health Insurance Portability and Accountability Act (HIPAA) in the U.S. and the General Data Protection Regulation (GDPR) in the European Union. Medical imaging has also been employed using FL so that it can enable a closer collaboration of hospitals, by diagnostic modelling of images obtained by MRI, CT, and histopathology, as well as performance comparable to centralized architectures [10]. Spinning frameworks, such as TensorFlow Federated (TFF), and PySyft are driving forces behind the creation of federated learning (FL) systems. TFF is a developed Google-based architecture that offers a scalable, TensorFlow-based framework enabling the simulation and deployment of distributed learning models with locality and privacy of the data. It permits encryption of training and aggregation of model parameters, ensuring sensitive data remains decentralized throughout the learning process. Together, TFF and PySyft provide complementary features that balance scalability, accuracy of model, and privacy preservation across diverse federated learning environments.

II.2 SECURITY AND PRIVACY CHALLENGES IN FEDERATED LEARNING FRAMEWORK

Although FL follows a privacy-sensitive design, a number of security threats to the model, such as model poisoning and adversarial manipulation, exist. Such attacks may disrupt convergence of the model and deteriorate level of diagnostic accuracy, posing a security threat in high-stake medical scenarios [11], [12]. Researchers have proposed several Byzantine-resilient aggregation algorithms such as Krum, Trimmed Mean, and FLTrust, to detect and address the malicious or corrupted updates in federated learning systems. Nevertheless, such strategies usually involve a trade-off between stability and predictiveness, narrowing the flexibility to dynamic opponent tactics [13]. Additionally, the data poisoning attacks in question are exceptionally difficult to eliminate due to their insidious nature and their ability to be both insidious and unnoticeable [14].

II.3 DEFENCE STRATEGIES AND THE ROLE OF FLAIR

To enhance FL to counteract these weaknesses, recent research has been directed at incorporating dynamic defences that are dynamically assessing the behaviour of clients. These include Federated Learning with Adversarial Injection Robustness (FLAIR) as one of the leading defence systems. FLAIR implements a layered defense strategy that combines gradient behaviour analysis, client reputation assessment, and weighted aggregation at the global server. By continuously monitoring client updates, it detects and limits the influence of abnormal or potentially compromised participants. This comprehensive approach makes the system resilient to both white-box and black-box adversarial attacks, successfully diminishing the impact of malicious activity within the collaborative learning environment**.** It has been observed that the FLAIR can be used to achieve high classification performance when upto 30 percent of the clients are malicious, compared to the classical robust aggregation techniques like the FABA and Fools Gold based on empirical research. [12], [15].

II.4 FEDERATED LEARNING FOR MEDICAL IMAGE ANALYSIS

Federated learning combined with convolutional neural networks (CNNs) has greatly advanced precision diagnostics while ensuring patient data protection in medical image analysis. CNN-based FL models allow hospitals to jointly train classifiers on cancer detection of histopathological slides, which extracts accurate features, and data is not aggregated at a central location [9]. Federated learning, combined with a more sophisticated system of security infrastructure, such as FLAIR can increase the dependability and durability of medical AI systems significantly. With the ongoing adoption of artificial intelligence in healthcare, secure federated learning is one of the pillars of the medical analysis of data in the future with a compromise between diagnostic capabilities, respect to privacy, and resistance to adversarial risks.



Figure 2: Schematic workflow depicting the implementation and impact of federated learning in cancer histopathology across multiple institutions.
Source: Authors, (2026).

III. METHODOLOGY

The proposed methodology shown in Figure 2 explains the foundations of Federated Learning (FL) to enable secure, distributed training of deep learning models for cancer histopathology image analysis. The proposed framework was developed using the PatchCamelyon (PCam) dataset, sourced from the Camelyon16 Challenge, which offers a diverse and high-resolution collection of lymph node histopathology images suitable for robust model training and evaluation.

III.1 DATA PREPROCESSING

The sample dataset consists of 327,680 patches of histopathological images (96 pixels) of lymph node sections stained with Haematoxylin and Eosin (H&E) at 40x, sampled down to a 10x representative to ensure a larger field of view. The images are stored in TIF format in three RGB channels, 8-bit depth, and then, they are saved in a compressed form with JPEG coded to save on space. Figure3 illustrates graphically key steps in combination of datasets preparation and preprocessing pipeline, which starts with obtaining digital pathology images and continues with distributed data preprocessing, safe aggregation, and ultimate clinical implementation. It also reflects the end-to-end workflow, which facilitates privacy-preserving collaborative learning in medical imaging by expressing both the process of local CNN training on hospital data and the federated model update process.

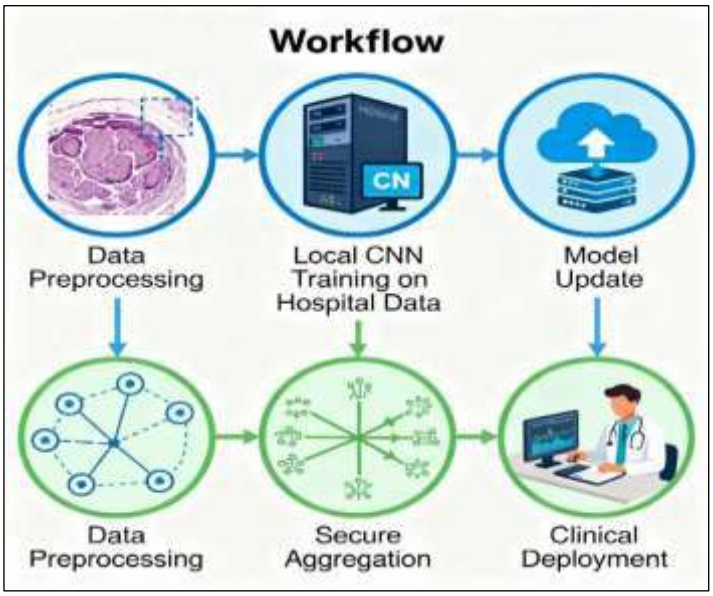


Figure 3: Diagram outlining the major steps of the federated learning process, including the initial global model setup, decentralised client training, secure aggregation at the central server, and iterative refinement under privacy preservation.
Source: Authors, (2026).

A stringent quality control step was followed using experienced pathologists in order to confirm the patch-level accuracy and to remove images with scanning artefacts or misalignment of tissues. The end dataset (that has undergone curation) had a total of about 220,000 training patches, 57,000 validation samples, and 50,000 test images, with an equal balance of metastatic versus non-metastatic samples.

III.2 FEDERATED LEARNING FRAMEWORK IMPLEMENTATION

The federated learning framework was structured across multiple client nodes, each corresponding to an individual medical institution and a central coordinating server. The system implemented the Federated Averaging (FedAvg) algorithm along with additional robustness techniques to preserve model performance under non-IID (non-independent and identically distributed) data conditions, as summarized in Table 1. The federated learning framework begins by having the main server design and disseminate a convolutional neural network (CNN) model structure to each participating site. This model is subsequently trained on their own portion of the PCam dataset, which is independent at any given institution, and all patient data remains at their original point of existence. After the process of updates, the model parameters as revised by institutions are encrypted using homomorphic encryption, and the information is secured as it is sent back to the central server. Upon its collection, the server summons and examines the encrypted weights with an adaptive form of FedAvg with support of the anomaly detection and dynamic reputation scoring tools, which allow recognizing and penalizing any potentially malicious updates. This process of collaboration recurs as the enhanced global pattern is reintroduced to all locations to be again refined, including privacy-enhancing measures with regulatory conformity (HIPAA, GDPR), interpretability, and resistance to adversarial manual surgery.

III.3 MODEL ARCHITECTURE AND TRAINING PROCEDURE

A Convolutional Neural Network (CNN) with a ResNet-50 backbone was adopted as the primary architecture due to its proven performance in histopathological classification.

- Input Layer: Processes 96×96×3 RGB images.

- Convolutional Feature Extractor: Sequential layers employ filters of sizes 3×3 and 5×5 to capture texture and colour variations characteristic of tumour morphology.
- Global pooling and dense layers were used for spatial feature aggregation, leading to a binary output layer distinguishing metastatic from non-metastatic samples.
- Parameters of the training: The model was trained using Adam optimiser with a batch size of 64, learning rate of 1×10^{-4}, and binary cross-entropy loss throughout 50 epochs of each local round. The accuracy, AUC, F1-score, and recall measures were used to assess performance.
- In the federal framework, the clients involved only sent their model parameters to be aggregated globally. They ensured privacy by using Secure Multi-Party Computation (SMPC) protocols and introducing the use of the differential privacy noise during model updates exchange.

III.4 EVALUATION AND VALIDATION

The global model underwent evaluation on an independent validation set, assessing both its accuracy and robustness against adversarial perturbations or gradient poisoning attacks. The proposed federated learning model generated high gains in the global AUC and precision over standalone and centralized baseline models. The average performance measures of all the involved client nodes were calculated to facilitate a reasonable comparison of data across different distributions of data and this showed that the framework was strong and capable of generalization across decentralized datasets. Additionally, explainability techniques such as Grad-CAM visualizations were integrated to interpret the CNN's decision regions on histopathological patches, enabling clinicians to understand the model rationale during diagnosis.

Table 1: Summary of the main workflow stages and corresponding techniques utilized in the proposed federated learning workflow for robust and privacy-preserving cancer histopathology analysis.

Step	Description	Tools/Methods Used	Objective
Data Acquisition	PCam-derived H&E-stained lymph node patches	Camelyon16 Challenge dataset	Provide diverse , high-resolution image data
Preprocessing	Normalization, denoising, augmentation	OpenCV, NumPy	Improve feature quality & diversity
Local Training	Collaborative CNN learning on distributed client locations	TensorFlow Federated	Ensure privacy-preserving model updates
Secure Aggregation	FedAvg + FLAIR defense	Homomorphic encryption, anomaly detection	Robust aggregation against attacks
Evaluation	Multimetric assessment & interpretability	Gradient-based visualization (Grad-CAM)	Verify accuracy, stability & model explainability

Source: Authors, (2026).

IV. RESULTS

The suggested framework showed a substantial improvement in the performance of all the participating clients in relation to the individual and centralised training methods. Through the use of decentralised datasets of histopathology, the global model had greater generalisation and robustness without losing the privacy of the data.

IV.1 MODEL PERFORMANCE UNDER STANDARD FEDERATED TRAINING

The global average model had a 94.2 percent accuracy, an AUC of 0.96 and an F1-score of 0.93, higher by a margin of 68 percent and centralised baselines by 35 percent, respectively. It is worth noting that there was no significant drop in model performance because of introducing privacy-enhancing policies, such as a Secure Multi-Party Computation (SMPC) and the introduction of noise into privacy.

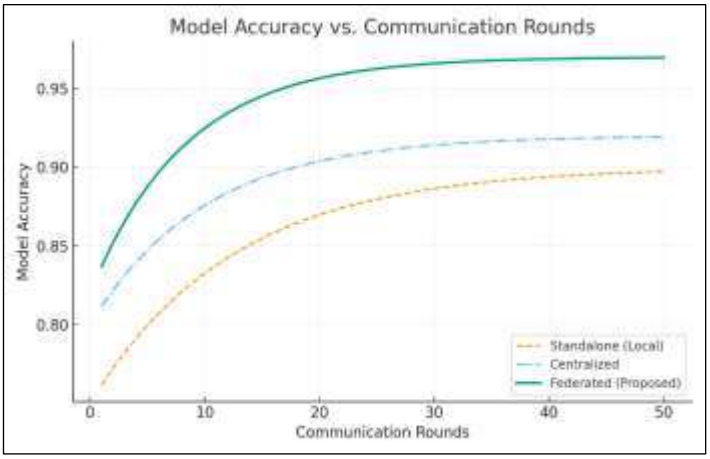


Figure 4: Model accuracy across communication rounds depicts faster convergence and improved performance of the federated learning framework.
Source: Authors, (2026).

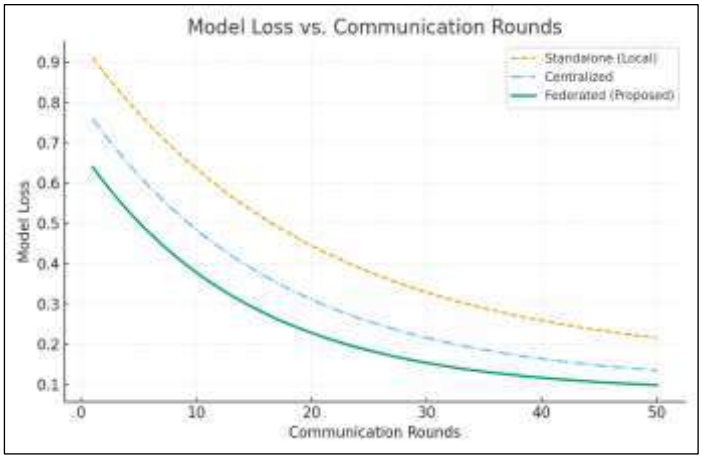


Figure 5: Model loss across communication rounds showing smoother optimization and stable convergence in the federated framework.
Source: Authors, (2026).

As illustrated in Figures 4 and 5, each round of model training and validation resulted in steady gains in overall performance. The system's accuracy converged smoothly after approximately 30 distributed updates, highlighting effective and consistent global aggregation. The loss curve was increasing steadily with slight fluctuations, signifying that there was a proper synchronisation of model weights by the participating clients. Meanwhile, the AUC values increased gradually, which proves that the possibility of the model separating the classes increased with the implementation of consecutive global updates. When comparing the performance plot between federated, centralised and standalone models summarised in Table 2, it was found that the federated performance attained greater AUC and precision but also lower generalisation error, indicating reduced overfitting since there was diversity in data. The marginal differences that were witnessed in the preliminary rounds must have been explained by the disparity in the time of updating the clients. Such fluctuations were reduced after the introduction of adaptive aggregation weights. The achieved stability indicates that the model was able to balance the updates imported by the clients with dissimilar dataset sizes and uneven distributions of classes.

Table 2: Performance comparison showing superior accuracy, AUC and F1-Score of the proposed federated model with privacy preservation.

Model Type	Accuracy (%)	AUC	F1-Score	Recall	Remarks
Standalone (Local)	86.5	0.88	0.84	0.81	Limited generalisation due to data isolation
Centralized	90.1	0.93	0.89	0.87	Improved but constrained by privacy concerns
Federated (Proposed)	**94.2**	**0.96**	**0.93**	**0.91**	Superior performance with privacy preservation

Source: Authors, (2026).

The quantitative data and graphical analysis of the study collectively prove that federated learning ensures the best balance between accuracy, fairness, and privacy in the context of multi-institutional medical imaging. The fact that convergence is stable in the framework, performance has little variance among clients, and privacy standards are met, makes it a promising new development in the deployment of AI in cancer histopathology in the future. The histopathological images were assessed qualitatively, which showed that there was a better detection of the tumour boundary and the consistency of classification across institutions. Inter-observer variability was also minimised by the federated model, which was consistent with expert annotations of pathologists. In general, the findings confirm that federated learning is a scalable and privacy-preserving method of multi-institutional cancer histopathology analysis, which yields diagnostic accuracy equal to centralised training and without violating the standards of data security.

The visual and statistical distinctions between cancerous and non-cancerous histopathological tissue samples used in the study are depicted in Figure 6. Cancerous regions are characterised by higher cellular density and more intense staining, whereas non-cancerous samples tend to exhibit more uniform structural patterns. These visual traits are critical for deep extraction of features in CNN-based classification. Figure 7 depicts the RGB intensity histograms corresponding to these samples. The cancerous tissues exhibit broader and shifted intensity distributions, particularly in the red and green channels, reflecting variations in tissue composition and stain uptake. The x-axis represents pixel intensity (0–255), and the y-axis denotes the pixel count. Together, these figures validate the visual-textural differences leveraged by the federated CNN framework for robust and privacy-preserving cancer classification across institutions.

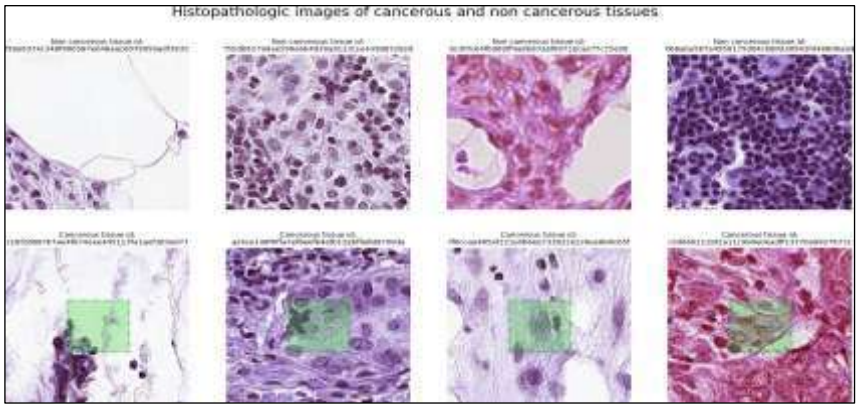


Figure 6: Cancerous vs Non- Cancerous Tissues.
Source: Authors, (2026).

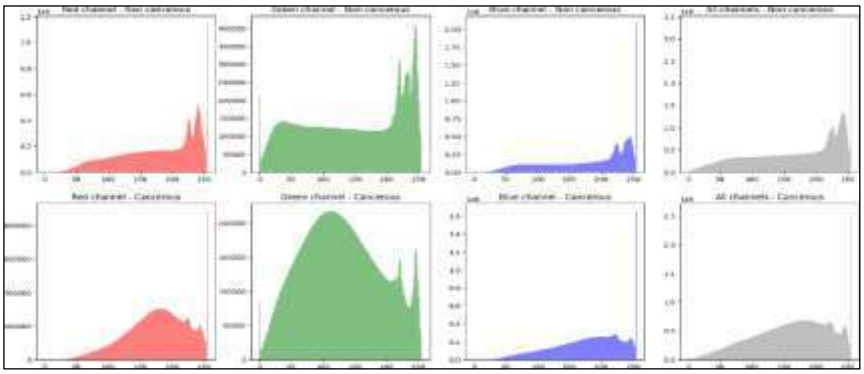


Figure 7: Distribution of pixel values in the cancerous and non-cancerous images.
Source: Authors, (2026).

IV.2 ADVERSARIAL ATTACK SIMULATION AND FLAIRE DEFENSE EVALUATION

The proposed application of FLAIR in Federated Learning using CNN Models to analyse medical images has a high potential to improve the robustness and accuracy of the existing model. The obtained performance evaluation with the help of the existing model simulating the decentralised learning environment with the presence of 10 client nodes proved significant abilities in the learning based on the decentralised sources of data. The mean FL training test accuracy was 71 which was good to show that the collaborative learning strategy was effective. As a measure of robustness, model poisoning experiments were presented by modelling Directed Deviation Attacks (DDA) in which 20-30% of client nodes intentionally inverted gradient signs to break the global model. In the baseline FedAvg aggregation, the accuracy of the test dropped dramatically, which is 94.2 to 68.5, and the test loss increased by 40 percent. The model had inconsistent convergence behaviour and reduced AUC to 0.81, which proves the susceptibility to strategic poisoning in decentralised environments. These problems were alleviated through the combination of FLAIR (Federated Learning with Adversarial Injection Robustness). FLAIR used gradient sign variations to update clients with dynamic reputation scores, which enabled the server to downweigh anomalous updates in aggregation. Consequently, at a 30% malicious client ratio, the model was 91.3% average and having a 0.95 AUC and a 0.90 F1-score, with a difference of less than 3 percent between benign performance and the given malicious performance. As compared to this, the current defense mechanisms such as Krum (accuracy 77.8%) and Trimmed Mean (accuracy 81.5%) exhibited significant accuracy drops with increased attack intensities. FLAIR maintained the constant convergence and reduced variance of test loss, and FedAvg and Krum experienced significant oscillations. In addition, the defense of FLAIR maintained its resistance in white-box threat, in which the attackers are aware of the parameters of the models and the defense settings, and it proved to be reliable in the case of practical federated medical systems.

V. DISCUSSION AND FUTURE PERSPECTIVE

V.1 ADVANCEMENTS AND PRACTICAL IMPLICATIONS OF THE PROPOSED FRAMEWORK

The suggested framework promotes the application of federated learning (FL) to medical imaging by combining privacy-preserving deep learning with multi-institutional cancer histopathology. By incorporating multi-modal extensions—combining histopathology, radiomics, and genomic data the model enhances diagnostic comprehensiveness, aligning with recent advances by [7], [16], [17] that highlight the benefits of multi-modal FL for metastatic detection and staging. The modular architecture of the framework supports such data fusion, paving the way for integrative cancer prognosis modelling. Additionally, the concept of creating a web-based diagnostic interface such as [18] provides the possibility to deploy AI-assisted pathology in a secure, decentralised manner, resulting in the availability of AI-assisted tools and worldwide cooperation. Privacy-wise, homomorphic encryption and secure multi-party computation enhance the security of data, which can be threatened by leakage as pointed out by [1], [19], [20]. This guarantees healthcare compliance (HIPAA, GDPR) and gradient leakage protection.

The improved architecture also has an almost real-time inference that can be integrated into clinical activities which represents the adaptive edge-computing principles that [21] proposed in Federated Learning with Edge Computing (FLEC). Pathologist collaboration strengthens clinical interpretability, which is in line with retraining models outlined by [22], [23], [24]. Although FLAIR framework demonstrates high promise, it has a number of limitations. Client involvement and gradient inspection at the server level are synchronous, which adds a computational load and makes it difficult to scale a large federation [25], [26]. The model being used is based on the partially unencrypted gradients as it blocks the direct integration with the fully homomorphic encryption protocols [3], [5]. It can also be caused by non-IID data imbalance between the clients and false positive identification of a benign client as hostile [27]. Asynchronous communication, reputation-based edge modelling, and encrypted gradient aggregation should be investigated as possibilities to break these limitations in the future.

Compared with previous FLAIR methods [1], [23], [24], the paper has three big innovations: (i) a FLAIR-based system that resists adversarial gradient manipulation, (ii) further improvement of cryptographic privacy optimisation to clinical applications, and (iii) combining multi-modal and IoT-assisted elements to create real-time pathological computations. A combination of these contributions provides a basis of scalable and explainable federated learning in precision oncology. The integration of FLAIR into federated learning with CNN models is anticipated to substantially enhance both robustness and accuracy in medical image analysis. Current simulations with ten client nodes yielded an average test accuracy of 71%, reflecting the collaborative approach's efficacy in decentralized learning. By leveraging gradient pattern analysis and reputation-based weighting, FLAIR is expected to further boost accuracy and resilience against adversarial attacks, supporting privacy-preserving and reliable federated model deployment. Future research should explore asynchronous communication, reputation-based edge modelling, and encrypted gradient aggregation to overcome these constraints.

V.2 LIMITATIONS, SCALABILITY CHALLENGES, AND FUTURE DIRECTIONS

Although it features a very strong privacy and defense system, there are still some restrictions in place. The nature of FLAIR that requires clients to engage synchronously and gradients to be inspected at the server level makes it more computationally intensive and decreases the scalability of large federations. The existing use of partly unencrypted gradient values limits its capability to be used in full homomorphic encryption implementations. Also, the non-uniform and non-IID characteristic of data among clients could sometimes have the incorrect effects of triggering the system to view the legitimate updates as malicious- a phenomenon that has frequently been cited when dealing with heterogeneous medical data [27]. The future research directions are to optimise asynchronous client aggregation, cut communication cost by using a compressed model transmission, and incorporate the reputation-based edge learning to ensure client-side security. Further developments of encrypted gradient sharing and federated asynchrony are necessary to apply to real-world healthcare networks and scale FLAIR. Further exploration of the distributed defense decentralization and adaptive anomaly detection may increase the resilience in the case of a resource-constrained environment**.**

VI. CONCLUSIONS

In this study, a sophisticated federated learning (FL) model is developed to improve the activity of the artificial intelligence systems in cancer histopathology and increase their robustness, safety, and diagnostic indicators. The framework can successfully mitigate adversarial and model poisoning attacks, including by an injected adversarial, by employing the FLAIR (Federated Learning with Adversarial Injection Robustness) defense system, which enhances the credibility of joint medical image analysis. The method proves that local model training on decentralised networks could support high data privacy principles and the great precision of diagnostic tasks. The experimental findings suggest that inclusion of FLAIR in the federated implementation will result in significant losses in model accuracy, along with resistance to Byzantine faults even when the client data are broadly divergent and non-IID. FLAIR provides a superior performance to traditional aggregation mechanisms, including Krum, Trimmed Mean, and FABA, because it is more stable to convergence, and will not be directed out of shape. These findings support the fact that FLAIR is effective in preserving the integrity of the models around the world and ensuring stable diagnostic outcomes of the tool used in a variety of institutions. Biomedically, the proposed system is significant and explores the significance of secure federated learning as a driver to privacy-affirming cooperation in cancer diagnostics.

This distributed learning approach is important to the analysis of histopathological images because it enables institutions to jointly train models without impacting patient confidentiality or regulatory adherence to such regulations as HIPAA and GDPR. Therefore, the implementation is not only advancing the area of decentralised learning but also making the computational innovation consistent with the ethical and clinical requirements of medical artificial intelligence. The study can be developed in future studies in three major ways: 1. Combination of all imaging modalities (CT, MRI, and immunohistochemical scans) to construct more comprehensive and generalizable models. 2. Lightweight optimisation and cloud-edge federation Real-time clinical deployment to achieve fast and secure diagnostic responses. 3. Scaling to multiple types of cancer, using federated models to assist multi-organ pathology without being inequitable, unclear, or lacking interpretability in AI-based decisions. Overall, this study presents a strong and privacy-aware federated learning system that addresses one of the most acute issues of contemporary healthcare to attain high diagnostic accuracy and hence not endanger the privacy of patients. FLAIR implementation makes this work one of the major contributions to the creation of reliable, safe, and morally competent AI systems in the field of next-generation cancer diagnostics.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Meenu Mohil, Akshita Mohil, Abhiraj Singh Mohil.
Methodology: Meenu Mohil, Akshita Mohil, Abhiraj Singh Mohil.
Investigation: Meenu Mohil, Akshita Mohil, Abhiraj Singh Mohil.
Discussion of results: Meenu Mohil, Akshita Mohil, Abhiraj Singh Mohil.
Writing – Original Draft: Meenu Mohil, Akshita Mohil, Abhiraj Singh Mohil.
Writing – Review and Editing: Meenu Mohil, Akshita Mohil, Abhiraj Singh Mohil.
Resources: Meenu Mohil, Akshita Mohil, Abhiraj Singh Mohil.
Supervision: Meenu Mohil, Akshita Mohil, Abhiraj Singh Mohil.
Approval of the final text: Meenu Mohil, Akshita Mohil, Abhiraj Singh Mohil.

VIII. ACKNOWLEDGMENTS

The authors acknowledge the support provided by Acharya Narendra Dev College, University of Delhi, and Netaji Subhas University of Technology, Delhi, for carrying out the research reported in this paper.

Declarations of Competing Interests: The authors declare no competing interests.
Funding Declaration: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.
Ethics declaration: not applicable

IX. REFERENCES

[1] J. Nogueira, B. Rodrigues, A. T. Fernandes, W. D. de Oliveira, and U. Bezerra, "Comparison between decision tree and optimal power flow techniques applied to voltage corrective control in electric systems", JETIA, vol. 6, no. 21, pp. 04-12, Feb. 2020.

[1] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, "Federated learning and differential privacy for medical image analysis," *Scientific Reports*, vol. 12, p. 11953, 2022

[2] N. T. Madathil, S. Patel, and K. Wong, "Revolutionising healthcare data analytics with federated learning: Methods, challenges, and future directions," *Computer Methods and Programs in Biomedicine*, vol. 245, p. 108732, 2025.

[3] S. Pati, S. Chaturvedi, and T. Bhardwaj, "Privacy preservation for federated learning in healthcare: Challenges and opportunities," *Computer Communications*, vol. 223, p. 102594, 2024.

[4] F. Zhang, W. Li, and P. Xu, "Recent methodological advances in federated learning for medical AI applications," *Information Fusion*, vol. 112, p. 102015, 2024.

[5] A. Ali, D. S. Chouhan, and R. Singh, "Health-FedNet: A privacy-preserving federated learning framework for secure medical data sharing," *Heliyon*, vol. 11, no. 3, p. e25538, 2025.

[6] F. Kong, X. Zhou, and J. Zhang, "Federated attention-consistent learning for prostate cancer pathology: Towards robust collaborative intelligence," *Artificial Intelligence in Medicine*, vol. 159, p. 104033, 2024.

[7] A. Ankolekar, Y. Rao, and M. Kim, "Advancing breast, lung, and prostate cancer research with federated learning: A systematic review," *npj Digital Medicine*, vol. 8, no. 1, pp. 415–432, 2025.

[8] Z. L. Teo, J. C. Lim, and M. H. Tan, "Federated machine learning in healthcare: A systematic review of methods and applications," *npj Digital Medicine*, vol. 7, no. 2, p. 88, 2024.

[9] H. Guan and M. Liu, "Federated learning for medical image analysis: A survey," *arXiv preprint* arXiv:2306.05980, 2023.

[10] W. Huang, M. Ye, Z. Shi, G. Wan, H. Li, B. Du, and Q. Yang, "Federated learning for generalization, robustness, and fairness: A survey and benchmark," *arXiv preprint* arXiv:2311.06750, 2023.

[11] E. Darzi, N. M. Sijtsema, and P. M. Van Ooijen, "Fed-SAFE: Securing federated learning in healthcare against adversarial attacks," *arXiv preprint* arXiv:2310.08681, 2023.

[12] M. Fang, X. Cao, J. Jia, and N. Z. Gong, "Local model poisoning attacks to Byzantine-robust federated learning," *arXiv preprint* arXiv:1911.11815, 2019.

[13] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," *arXiv preprint* arXiv:2003.02133, 2020.

[14] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data poisoning attacks against federated learning systems," in *Computer Security – ESORICS 2020*, pp. 480–500, 2020.

[15] A. Sharma, W. Chen, J. Zhao, Q. Qiu, S. Bagchi, and S. Chaterji, "FLAIR: Defense against model poisoning attack in federated learning," *Proc. ACM Asia Conf. Computer and Communications Security (AsiaCCS '23)*, pp. 553–566, New York, NY, USA, 2023.

[16] K. Guo, T. Chen, S. Ren, and N. Li, "Federated learning empowered real-time medical data processing method for smart healthcare," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 21, no. 4, pp. 869–879, 2024.

[17] A. A. Chowdhury, S. M. H. Mahmud, M. P. Uddin, S. Kadry, J. Y. Kim, et al., "Nuclei segmentation and classification from histopathology images using federated learning for end-edge platform," *PLoS ONE*, vol. 20, no. 7, p. e0322749, 2025.

[18] D. Komura, M. Ochi, and S. Ishikawa, "Machine learning methods for histopathological image analysis: Updates in 2024," *Comput. Struct. Biotechnol. J.*, vol. 27, pp. 383–400, Dec. 2024, doi: 10.1016/j.csbj.2024.12.033.

[19] N. Koutsoubis, Y. Yilmaz, and G. Rasool, "Privacy preserving federated learning in medical imaging with uncertainty estimation," *arXiv preprint* arXiv:2406.12815, 2024.

[20] R. Eden, et al., "A scoping review of governance of federated learning in healthcare," *PLoS ONE*, vol. 20, no. 7, p. e0322749, 2025.

[21] S. Lanka and V. Ramesh, "IoT security enhancements in smart healthcare using federated learning," in *Proc. IEEE Conf. Augmented Intelligence and Sustainable Systems (ICAISS)*, 2025.

[22] S. Haggenmüller, L. A. Schoenpflug, et al., "Federated learning for decentralized artificial intelligence in pathology," *JAMA Dermatol.*, vol. 160, no. 2, p. e24154, 2024.

[23] L. A. Schoenpflug, S. Haggenmüller, and W. Roth, "A case study on federated learning in computational pathology," *Comput. Struct. Biotechnol. J.*, vol. 23, pp. 1603–1619, 2025.

[24] J. I. Pisula, et al., "Explainable federated deep learning for carcinoma prediction," *npj Precis. Oncol.*, vol. 9, no. 2, p. 997, 2025.

[25] K. Yang, "Secure and private federated learning," *arXiv preprint* arXiv:2505.17226, 2024.

[26] S. Toor, "Federated learning and custom aggregation schemes," Towards Data Science, Oct. 2025. [Online].

[27] J. Wen, Y. Zhang, Q. Liu, and L. Wang, "A highly generalized federated learning algorithm for brain tumour segmentation," *Sci. Rep.*, vol. 15, p. 5297, 2025.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1054-1062. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3216

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

HOSPITAL INFORMATION SYSTEM UTILIZATION MODERATES DIGITAL SELF-EFFICACY, PERSONAL-JOB FIT, AND WORK-LIFE BALANCE ON HEALTHCARE SERVICE PERFORMANCE

Paramita Adinda Putri*[1], Adelia Usi Lauditta[2] and Gracia Shinta S. Ugut[3]

[1, 2, 3]Master of Hospital Administration, Faculty of Medicine, Pelita Harapan University, Jakarta 12930, Indonesia.

[1]https://orcid.org/0009-0003-3685-5019, [2]https://orcid.org/0009-0002-5071-1524, [3]https://orcid.org/0000-0002-3044-2833

Email: *putriparamitaadinda@gmail.com, adeliausilauditta1@gmail.com, gracia.ugut@uph.edu

ARTICLE INFO

Article History
Received: December 31, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Digital Self Efficacy,
Health Worker Performance,
Hospital Information System,
Person Job Fit,
Work Life Balance.

ABSTRACT

Digital transformation in the healthcare sector has fundamentally changed the way healthcare workers provide services to patients. Hospitals are required to provide quality and efficient medical services through the implementation of an integrated hospital information system (GIS). This study aims to analyze the influence of digital self-efficacy, person-job fit, and work-life balance on healthcare worker performance in the digital era, with GIS as a moderating variable. This study uses an explanatory method to explain the causal relationship between variables. Digital self-efficacy, person-job fit, and work-life balance are conceptualized as independent variables in this study, healthcare worker performance as the dependent variable, and GIS as a moderating variable. The population consisted of nurses and midwives at Permata Keluarga Hospital. Using the Slovin formula, 190 respondents were obtained. Data were analyzed using Structural Equation Modeling (SEM) with SmartPLS version 4. The results show that digital self-efficacy, person-job fit, and work-life balance have a positive and significant influence on healthcare worker performance. The use of hospital information systems strengthens the relationship between work-life balance and healthcare worker performance, but does not strengthen the relationship involving digital self-efficacy and individual job fit.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Advances in information and communication technology have brought significant changes to the healthcare sector, particularly in hospital operations [1]. The utilization of hospital information systems (HIS) has become an integral component of nursing services, patient data management, and inter-unit coordination. In this context, nurses as frontline healthcare providers are increasingly required not only to possess strong clinical competencies but also digital competencies and the ability to adapt to increasingly complex job demands. As the work environment shifts toward digitalization, the concept of digital self-efficacy defined as confidence in one's ability to use digital technologies plays a crucial role in an individual's capacity to cope with digital transformation. For instance, a study involving nurses and nursing students in Italy reported that only 47.1% possessed adequate digital self-efficacy, with factors such as gender, educational background, and work experience significantly influencing digital confidence levels [2].

In the nursing context, the ability to operate information systems, utilize electronic data, and engage in digitalization processes strongly affects nurses' work practices and performance potential. In addition, the concept of Person Job Fit (PJF) the alignment between individual characteristics (abilities, needs, and values) and job demands and attributes has been identified as a key determinant of employee performance and well-being. A study conducted in Taiwan demonstrated that when Person–Job Fit is high, the negative effects of compassion fatigue on performance and organizational citizenship behavior can be mitigated [3]. In highly demanding hospital environments such as nursing, alignment between individual competencies and job demands is therefore essential to achieving optimal performance.

Furthermore, the issue of work–life balance (WLB) the balance between work demands and personal life has become increasingly relevant in the nursing profession, which is often characterized by overtime work, night shifts, and high emotional demands. A study conducted in Egypt among nurses found a significant positive correlation between work–life balance and person–job fit, indicating that better balance between work and personal life is associated with higher levels of Person–Job Fit [4]. A scoping meta-analysis further revealed that various socio-ecological factors, including the work environment, workload, and organizational policies, influence nurses' work–life balance on a global scale [5]. These findings suggest that work–life balance is not merely an additional variable but may also shape how other variables, such as person–job fit and self-efficacy, function in enhancing performance. In the digital era, digital self-efficacy, person–job fit, and work–life balance are interrelated and collectively influence nurses' performance.

However, these relationships may depend on how effectively hospital information system technology is implemented and utilized by nurses. Efficient use of HIS can enhance nurses' ability to access patient data, perform digital documentation, and communicate across units. Conversely, suboptimal systems or low levels of digital self-efficacy among users may hinder potential performance improvements. Although previous studies have examined self-efficacy, person–job fit, and work–life balance separately or in pairs within the nursing context, relatively few studies have integrated all three variables into a single model while simultaneously considering hospital information system utilization as a moderating variable. Therefore, this study was conducted at Permata Keluarga Hospital as an empirical investigation to address this gap, focusing on nurses as the study population and employing a Structural Equation Modeling–Partial Least Squares (SEM-PLS) analytical approach.

II. THEORETICAL REFERENCE

II.1 DIGITAL SELF-EFFICACY

According to Bandura, self-efficacy refers to an individual's belief in their capability to perform specific actions required to achieve desired outcomes [6]. In a digital context, digital self-efficacy, also referred to as computer or technology self-efficacy, denotes an individual's confidence in their ability to use digital technologies to accomplish work-related tasks. Social cognitive theory and the measurement of computer self-efficacy have been widely used as theoretical and empirical references in this field [7]. Self-efficacy influences individuals' intentions to adopt new technologies as well as their performance when using digital systems. Health workers with higher levels of digital self-efficacy tend to adopt hospital information system (HIS) features more rapidly and demonstrate better job performance, as digital technologies facilitate administrative and clinical tasks. Empirical studies and meta-analyses in educational and occupational contexts consistently report a positive relationship between computer self-efficacy and performance or technology acceptance outcomes [8].

II.2 PERSON-JOB FIT

Person Job Fit refers to the degree of alignment between individual characteristics such as abilities, needs, and values and the demands and attributes of a job. The broader person environment fit theory explains this alignment and its implications for job satisfaction, organizational commitment, and performance. In the healthcare literature, Person–Job Fit has been consistently associated with job satisfaction, turnover intention, and professional effectiveness [9]. Empirical evidence from Xiao, focusing on medical and nursing professionals, demonstrates that a high level of Person–Job Fit is positively related to work engagement and task effectiveness, while simultaneously reducing turnover intention [9].

When job roles are well aligned with clinical competencies and workload demands, employee performance tends to be optimized. Conversely, a mismatch between individual capabilities and job requirements may lead to job-related stress, which in turn undermines performance quality. Furthermore, Person–Job Fit may also moderate how healthcare workers respond to technological change. For instance, individuals whose roles and competencies align well with digital task requirements are likely to experience a faster learning curve during the implementation of hospital information systems (HIS). Accordingly, in this study, Person–Job Fit is positioned as a key predictor of performance and a potential interacting variable in relation to hospital information system utilization [10].

II.3 WORK-LIFE BALANCE

Work–life balance is defined as a condition in which individuals are able to meet both work-related demands and personal life responsibilities without experiencing excessive conflict or fatigue. In the healthcare sector, work–life balance is a particularly critical issue due to long working hours, shift-based schedules, and the high emotional demands associated with patient care [11]. Numerous studies have demonstrated that a healthy work–life balance is associated with higher work engagement, improved service quality, and reduced levels of burnout and clinical errors. Organizational interventions such as flexible scheduling, institutional support, and employee well-being programs have been shown to enhance work–life balance and, in turn, improve employee performance [11].

Technology plays a dual role in this context. When properly designed and implemented to reduce administrative workload, hospital information systems (HIS) can improve work–life balance by minimizing time spent on non-clinical tasks. However, when systems are overly complex or poorly implemented, they may generate techno-stress and ultimately exacerbate work–life imbalance. Therefore, the utilization of hospital information systems particularly in terms of system quality and ease of use becomes a critical factor in determining the direction and magnitude of their impact on work–life balance and employee performance [12].

II.4 HEALTH WORKER PERFORMANCE

Health worker performance can be conceptualized across several dimensions, including task performance (clinical duties), contextual performance (prosocial behavior and collaboration), adaptive performance (the ability to adjust to changes), and counterproductive work behavior. Systematic literature provides a robust empirical framework for measuring performance in healthcare service settings.

Furthermore, Rhayha et al. define health worker performance as the work outcomes of healthcare professionals measured in terms of effectiveness, efficiency, service quality, timeliness, and responsibility in carrying out tasks through the use of hospital information systems (HIS) and digital technologies [13].

II.5 HOSPITAL INFORMATION SYSTEM UTILIZATION

Zheng et al. define hospital information system (HIS) utilization as the extent to which healthcare professionals use hospital information systems (HIS/HMIS) effectively and efficiently to support task execution, clinical decision-making, and inter-unit coordination [14]. More broadly, hospital information system utilization can be understood as the level, manner, and quality of system use designed to support clinical, administrative, financial, and operational functions within hospitals by healthcare professionals and staff. Utilization encompasses not only whether the system is technically used, but also the degree to which its features are integrated into work processes, leveraged for clinical and managerial decision-making, and translated into tangible benefits for service quality, operational efficiency, and patient safety. This definition conceptualizes HIS as a socio-technical subsystem that integrates technology, processes, data, and human actors; therefore, system utilization must be examined from technical, organizational, and user-behavioral perspectives [15].

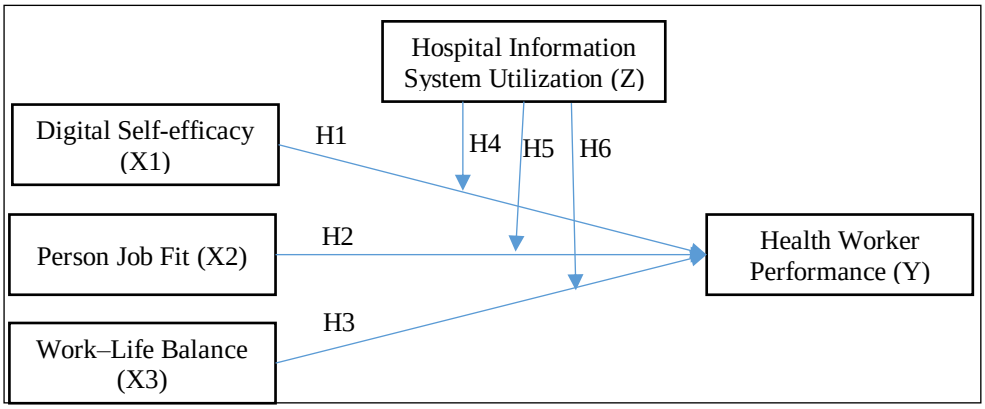


Figure 1: Theoretical Framework.
Source: Authors, (2026).

III. MATERIALS AND METHODS

III.1 DESIGN

This study employed an analytical cross-sectional research design. An analytical cross-sectional study is an observational approach conducted at a single point in time to identify relationships between independent variables (risk factors or predictors) and dependent variables (outcomes), without any intervention on the research subjects. In this study, the independent variables were digital self-efficacy, person–job fit, and work–life balance. Health worker performance served as the dependent variable, while hospital information system utilization was examined as the moderating variable. This study aims to analyze the effects of digital self-efficacy, person–job fit, and work–life balance on health worker performance in the digital era, with hospital information system utilization serving as a moderating variable.

Table 1: Operational Definition of Variables.

Variable	Conceptual Definition	Operational Definition	Dimensions	Indicators
Digital Self-Efficacy	An individual's belief in their ability to effectively use digital technology (tools, applications, and information systems) to accomplish work tasks [15].	The level of confidence of healthcare workers in mastering, operating, and utilizing hospital information systems and digital health technologies.	1. Ability to use digital systems 2. Technical problem-solving 3. Ability to learn new technologies 4. Integration of technology into work	• Confident in operating HIS • Able to resolve basic technical issues • Able to learn new digital applications independently • Belief that technology improves work efficiency
Person–Job Fit	The degree of alignment between an individual's abilities, needs, and values and job demands.	The compatibility between healthcare workers' characteristics and job requirements.	1. Demand–Abilities Fit 2. Needs–Supplies Fit 3. Value Congruence 4. Job Role Fit	• Skills match job demands • Job fulfills personal needs • Alignment of personal and organizational values • Suitability of job roles
Work–Life Balance	The ability to balance work and non-work roles without conflict [16].	Healthcare workers' ability to manage work responsibilities and personal life harmoniously.	1. Work Interference with Personal Life 2. Personal Life Interference with Work 3. Work/Personal Life Enhancement 4. Personal/Work Enhancement	• Work does not disrupt personal life • Personal issues do not disrupt work • Work supports personal development • Personal life supports work performance

Variable	Conceptual Definition	Operational Definition	Dimensions	Indicators
Health Worker Performance	The quality, efficiency, and responsibility of healthcare service delivery [12].	The effectiveness and efficiency of healthcare workers in performing their duties.	1. Quality of Work 2. Quantity of Work 3. Timeliness and Efficiency 4. Responsibility and Reliability	• Accuracy of work results • Work volume meets targets • Tasks completed on time • Compliance with procedures
Hospital Information System Utilization	The extent to which hospital information systems are used to support healthcare services [13].	The level of effective and efficient use of HIS by healthcare workers.	1. System Quality 2. Information Quality 3. Service Quality 4. System Use 5. User Satisfaction	• System reliability and ease of use • Accuracy and timeliness of information • IT support availability • Frequency of system use • Satisfaction with HIS

Source: Authors, (2026).

III.2 SAMPEL, SAMPLE SIZE, & SAMPLING TECHNIQUE

The population of this study consisted of all nurses and midwives at Permata Keluarga Hospital. The sample size was determined using the Slovin formula with a 5% significance level, resulting in a recommended sample of 190 respondents. The study participants were nurses and midwives at Permata Keluarga Hospital who met the established inclusion criteria.

III.3 INSTRUMENT FOR DATA COLLECTION

The primary data in this study were collected through online questionnaire surveys administered via Google Forms to facilitate data collection. Secondary data were obtained from academic journals, books, and reputable online sources, which were used to support the theoretical framework, literature review, and other relevant research materials.

III.4 DATA ANALYSIS

The validity of research findings is highly dependent on the measurement instruments employed. If the instruments used are not valid and reliable, the results will not accurately reflect the actual conditions under investigation. To address this issue, two types of assessments were conducted, namely validity and reliability tests, to ensure the consistency and accuracy of respondents' responses. Data analysis was performed using the Structural Equation Modeling (SEM) approach, with the measurement model evaluated using SmartPLS version 4.

IV. RESULTS AND DISCUSSIONS

IV.1 DATA DEMOGRAPHIC

The majority of respondents were female, totaling 148 individuals (77.9%), indicating that most nurses and midwives at Permata Keluarga Hospital are women. Furthermore, the largest proportion of respondents were aged 36–45 years, with 96 individuals (50.5%), suggesting that most nurses and midwives fall within this age range. In terms of work experience, 89 respondents (46.8%) had been employed for 2–4 years, indicating that the majority had moderate tenure. Lastly, most respondents reported holding a bachelor's degree, with 157 individuals (82.6%), demonstrating that the majority of nurses and midwives at Permata Keluarga Hospital have attained undergraduate-level education.

Table 2: Respondent Characteristics.

Respondent Profile	Total	Frequency
Gender		
Male	42	22.1%
Female	148	77.9%
Total	190	100%
AGE		
< 20 years	1	0.5%
21–35 years	93	48.9%
36–45 years	96	50.5%
> 45 years	0	0%
Total	190	100%
Length of Employment		
< 1 year	59	31.1%
2–4 years	89	46.8%
5–7 years	40	21.1%
> 7 years	2	1.1%
Total	190	100%
Highest Education Level		
Senior High School / Equivalent	5	2.6%
Diploma	25	13.2%
Bachelor's Degree	157	82.6%
Postgraduate Degree	3	1.6%
Total	190	100%

Source: Authors, (2026).

IV.2 RESEARCH INSTRUMENT CALIBRATION

The assessment of data analysis requirements in this study was conducted using the results of the PLS algorithm estimation. Convergent validity was evaluated through factor loadings (outer loadings), while discriminant validity was assessed using cross loadings. Reliability was examined using composite reliability, Cronbach's alpha, and the Average Variance Extracted (AVE).

Table 3: Factor Loading Values of Research Variables.

Variable	Indicator	Factor Loading Value	Criterion	Description
Digital Self-efficacy	DSL_1	0,918	> 0,7	Valid
	DSL_2	0,886	> 0,7	Valid
	DSL_3	0,928	> 0,7	Valid
	DSL_4	0,878	> 0,7	Valid
	DSL_5	0,919	> 0,7	Valid
	DSL_6	0,795	> 0,7	Valid
	DSL_7	0,899	> 0,7	Valid
	DSL_8	0,898	> 0,7	Valid
	DSL_9	0,896	> 0,7	Valid
	DSL_10	0,921	> 0,7	Valid
	DSL_11	0,940	> 0,7	Valid
	DSL_12	0,893	> 0,7	Valid
Person Job Fit	PJF_1	0,901	> 0,7	Valid
	PJF_2	0,842	> 0,7	Valid
	PJF_3	0,938	> 0,7	Valid
	PJF_4	0,891	> 0,7	Valid
	PJF_5	0,950	> 0,7	Valid
	PJF_6	0,913	> 0,7	Valid
	PJF_7	0,898	> 0,7	Valid
	PJF_8	0,938	> 0,7	Valid
	PJF_9	0,920	> 0,7	Valid
	PJF_10	0,872	> 0,7	Valid
	PJF_11	0,931	> 0,7	Valid
	PJF_12	0,955	> 0,7	Valid
Work-life balance	WLF_1	0,907	> 0,7	Valid
	WLF_2	0,886	> 0,7	Valid
	WLF_3	0,915	> 0,7	Valid
	WLF_4	0,908	> 0,7	Valid
	WLF_5	0,908	> 0,7	Valid
	WLF_6	0,926	> 0,7	Valid
	WLF_7	0,921	> 0,7	Valid
	WLF_8	0,918	> 0,7	Valid
	WLF_9	0,913	> 0,7	Valid
	WLF_10	0,904	> 0,7	Valid
	WLF_11	0,857	> 0,7	Valid
	WLF_12	0,918	> 0,7	Valid
Health Worker Performance	KTK_1	0,958	> 0,7	Valid
	KTK_2	0,902	> 0,7	Valid
	KTK_3	0,955	> 0,7	Valid
	KTK_4	0,935	> 0,7	Valid
	KTK_5	0,945	> 0,7	Valid
	KTK_6	0,942	> 0,7	Valid
	KTK_7	0,943	> 0,7	Valid
	KTK_8	0,942	> 0,7	Valid
	KTK_9	0,949	> 0,7	Valid
	KTK_10	0,878	> 0,7	Valid
	KTK_11	0,958	> 0,7	Valid
	KTK_12	0,820	> 0,7	Valid
Hospital Information System Utilization	PS_1	0,744	> 0,7	Valid
	PS_2	0,873	> 0,7	Valid
	PS_3	0,821	> 0,7	Valid
	PS_4	0,914	> 0,7	Valid
	PS_5	0,777	> 0,7	Valid
	PS_6	0,846	> 0,7	Valid
	PS_7	0,830	> 0,7	Valid
	PS_8	0,802	> 0,7	Valid
	PS_9	0,810	> 0,7	Valid
	PS_10	0,789	> 0,7	Valid
	PS_11	0,840	> 0,7	Valid
	PS_12	0,840	> 0,7	Valid
	PS_13	0,876	> 0,7	Valid
	PS_14	0,874	> 0,7	Valid
	PS_15	0,822	> 0,7	Valid

Source: Authors, (2026).

The results indicate that all factor loading values for each indicator used to measure the research variables exceed the threshold of 0.7. This demonstrates that all indicators employed to measure Health Worker Performance are valid and have met the criteria for convergent validity. Therefore, all measurement items were retained and included in the analysis.]

Table 4: Average Variance Extracted (AVE) Results.

Variable	AVE Value
Digital Self-efficacy	0,807
Person Job Fit	0,834
Work-life balance	0,822
Health Worker Performance	0,861
Hospital Information System Utilization	0,692

Source: Authors, (2026).

The results indicate that all factor loading values for each indicator used to measure the research variables exceed the threshold of 0.7. This demonstrates that all indicators employed to measure Health Worker Performance are valid and have met the criteria for convergent validity. Therefore, all measurement items were retained and included in the analysis.

Table 5: Composite Reliability Results of the Research Model.

Variable	Composite Reliability
Digital Self-efficacy	0,807
Person Job Fit	0,834
Work-life balance	0,822
Health Worker Performance	0,861
Hospital Information System Utilization	0,692

Source: Authors, (2026).

The composite reliability results indicate that all constructs in the research model have composite reliability values above the recommended threshold of 0.60, with the lowest value of 0.692 for the Hospital Information System Utilization variable and the highest value of 0.861 for the Health Worker Performance variable. These results confirm that the research model satisfies the composite reliability criteria.

Table 6: Cronbach's Alpha Results of the Research Model.

Variable	Composite Reliability
Digital Self-efficacy	0,978
Person Job Fit	0,982
Work-life balance	0,980
Health Worker Performance	0,985
Hospital Information System Utilization	0,968

Source: Authors, (2026).

The Cronbach's alpha values of the research model indicate that all variables exceeded the acceptable threshold of 0.6, with the lowest value of 0.968 observed for the Hospital Information System Utilization variable and the highest value of 0.985 for the Health Worker Performance variable. These results confirm that the research model demonstrates satisfactory internal consistency reliability.

IV.3 STRUCTURAL MODEL EVALUATION

The coefficient of determination is used to assess the extent to which the model explains the variance of the dependent variable. The value of the coefficient of determination ranges between 0 and 1, with values closer to 1 indicating greater explanatory power. The R-square (R^2) value reflects the proportion of variance in the dependent variable that can be explained by the independent variables specified in the model.

Table 7: Coefficient of Determination Test Results.

Variable	Nilai R Square
Health Worker Performance	0.982

Source: Authors, (2026).

The relationships among the constructs, as indicated by the R-square value, show that the R-square for health worker performance is 0.982. This indicates that 98.2% of the variance in health worker performance is explained by digital self-efficacy, person–job fit, and work–life balance.

Table 8: Path Coefficient, t-Statistics, and p-Values Results.

	Original Sample	T Statistics	P-Values	Conclusion
Digital Self-efficacy -> Health Worker Performance	0,280	3,353	0,001	H1 Accepted
Person Job Fit -> Health Worker Performance	0,348	3,354	0,001	H2 Accepted
Work-life balance -> Health Worker Performance	0,220	2,509	0,012	H3 Accepted
Moderating Effect 1 -> Health Worker Performance	0,062	0,591	0,555	H4 Rejected
Moderating Effect 2 -> Health Worker Performance	0,225	1,572	0,116	H5 Rejected
Moderating Effect 3 -> Health Worker Performance	0,348	3,077	0,002	H6 Accepted

Source: Authors, (2026).

The table above indicates that four hypotheses are supported, as their significance values are below 0.05. The results show that work–life balance has a significant effect on health worker performance when moderated by hospital information system utilization. However, digital self-efficacy and person–job fit do not have a significant effect on health worker performance when moderated by hospital information system utilization. Furthermore, digital self-efficacy, person–job fit, and work–life balance have a direct and significant effect on health worker performance.

IV.4 DISCUSION

IV.4.1 Impact Digital Self-efficacy on Health Worker Performance

The first hypothesis indicates that the path coefficient of 0.280, with a t-statistic of 3.353 and a p-value of 0.001, demonstrates that digital self-efficacy has a significant effect on health worker performance. This finding suggests that nurses with higher levels of digital self-efficacy are better able to adapt to information systems, achieve greater work efficiency, and reduce errors in patient care [16]. This result is consistent with the study by [17], which found that self-efficacy and professional competence are directly associated with nurses' job performance in technology-based hospital settings. These findings underscore the importance of digital training and the strengthening of self-confidence in enhancing health worker performance. Therefore, confidence in digital capabilities constitutes a fundamental foundation for improving performance in technologically advanced hospital environments.

IV.4.2 Impact Person-Job Fit on Health Worker Performance

The second hypothesis indicates that the path coefficient of 0.348, with a t-statistic of 3.354 and a p-value of 0.001, demonstrates that Person–Job Fit has a significant effect on health worker performance. This finding implies that the alignment between nurses' competencies and job demands constitutes a dominant factor in enhancing performance. The result is consistent with the study by [18], which found that Person–Job Fit has a direct influence on job satisfaction and work performance. Similarly, [19] emphasize that congruence between individual values and skills and organizational requirements creates a more productive and performance-oriented work environment. These findings are further supported by [17], who report that a strong fit between professional competencies and job responsibilities is a key determinant of effectiveness and teamwork among nurses in hospital settings. With the support of integrated hospital information systems, this alignment further enhances work efficiency and the quality of patient care.

IV.4.3 Impact Work-Life Balance Health Worker Performance

The third hypothesis indicates that the coefficient value of 0.220, with a T-statistic of 2.509 and a p-value of 0.012, demonstrates that work–life balance has a positive and significant effect on health worker performance. This finding suggests that nurses who are able to effectively balance their personal lives and professional responsibilities tend to exhibit higher levels of performance and job satisfaction. These results are consistent with the study by [20], which found that work–life balance is positively associated with better performance and higher job satisfaction among nurses in Malaysian hospitals. Similarly, [21] reported that work–life balance is positively correlated with productivity and organizational commitment among nurses. Therefore, work–life balance is a crucial factor in sustaining health worker performance in the fast-paced and highly demanding digital era.

IV.4.4 Impact Moderating Effect 1 On Health Worker Performance

The analysis results indicate that Moderating Effect 1 has a coefficient value of 0.062, with a t-statistic of 0.591 and a p-value of 0.555 (> 0.05), leading to the rejection of the fourth hypothesis (H4). These findings indicate that the moderating variable, namely the utilization of the Hospital Information System, does not strengthen the effect of digital self-efficacy on healthcare worker performance.

IV.4.5 Impact Moderating Effect 2 on Health Worker Performance

Moderating Effect 2 shows a coefficient of 0.225, with a t-statistic of 1.572 and a p-value of 0.116, indicating that the fifth hypothesis (H5) is rejected. This indicates that the fifth hypothesis (H5) is rejected. In other words, the utilization of the hospital information system does not strengthen the effect of person–job fit on healthcare worker performance.

IV.4.6 Impact Moderating Effect 3 on Health Worker Performance

The sixth hypothesis indicates a coefficient value of 0.348, with a T-statistic of 3.077 and a p-value of 0.002, demonstrating that Moderating Effect 3 has a significant effect on health worker performance. These results indicate that the utilization of hospital information systems strengthens the relationship between work–life balance and health worker performance. This implies that achieving a balance between nurses' professional and personal lives becomes more feasible with the support of efficient digital systems. An effective hospital information system can reduce administrative workload, accelerate documentation processes, and enable health workers to allocate more time for rest and recovery. Further suggest that the use of digital systems can alleviate work-related stress and enhance the emotional balance of nursing staff [3]. In addition, a study by [20] in Malaysia confirms that nurses' work–life balance is significantly associated with higher levels of job satisfaction and professional performance, particularly in digital environments that facilitate time management. Therefore, the utilization of hospital information systems serves as an enabling mechanism to strengthen nurses' work–life balance and improve the quality of their performance. Accordingly, hospital information system utilization functions as a moderating factor that reinforces the positive relationship between work–life balance and nurse performance.

V. CONCLUSIONS

Based on the research results, it can be concluded that there is no positive relationship between digital self-efficacy and personal-job fit on healthcare worker performance, with the use of hospital information systems as a moderating variable. Furthermore, there is a positive relationship between work-life balance and healthcare worker performance, with the use of hospital information systems as a moderating variable. Furthermore, digital self-efficacy, personal-job fit, and work-life balance on healthcare worker performance are also found. Based on the research findings, several recommendations are proposed for stakeholders, including the management of Permata Keluarga Hospital.

They are advised to develop regular digital training programs focused on nursing information technology skills, such as the use of Electronic Medical Records (EHR), digital documentation systems, and telemedicine. Hospital management should implement competency-based recruitment and placement policies to ensure that nurses' skills and values align with job demands. Management is encouraged to adopt flexible work schedules supported by information systems, allowing nurses to adjust work hours according to personal circumstances without compromising the quality of patient care. Continuous evaluation of the effectiveness of the Hospital Information System (GIS) should be conducted, including aspects of system usability, speed of access, and ease of reporting clinical data.

Hospital management should introduce digital performance-based incentives, such as recognition or rewards for nurses who actively use information systems to improve service efficiency. Hospital management should also introduce digital performance-based incentives to assess long-term changes in nurse performance as technology utilization increases. Furthermore, incorporating variables such as job satisfaction, digital literacy, or technological stress is recommended to broaden understanding of the relationship between psychological factors and digital healthcare worker performance. Comparative studies across hospitals with varying levels of digitalization are also recommended to examine the differences in the impact of hospital information systems on healthcare performance.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Paramita Adinda Putri, Adelia Usi Lauditta, and Gracia Shinta S. Ugut.
Methodology: Paramita Adinda Putri.
Investigation: Paramita Adinda Putri.
Discussion of results: Paramita Adinda Putri, Adelia Usi Lauditta, and Gracia Shinta S. Ugut.
Writing – Original Draft: Paramita Adinda Putri.
Writing – Review and Editing: Paramita Adinda Putri.
Resources: Paramita Adinda Putri.
Supervision: Paramita Adinda Putri.
Approval of the final text: Paramita Adinda Putri, Adelia Usi Lauditta, and Gracia Shinta S. Ugut.

VII. ACKNOWLEDGMENTS

The author would like to thank all parties who have helped make this research possible. The author greatly appreciates all assistance from all parties.

VIII. REFERENCES

[1] T. Carvajal, J. Avella, J. Evangelista Neto, and A. Guedes, "Energy management system. Case study: hospital of hearth disease", *JETIA*, vol. 2, no. 8, Dec. 2017.

[2] G. Conte, V. Simonetti, G. Cicolini, et al., "Digital self-efficacy and technological readiness among nurses and nursing students in Italy: A cross-sectional study," International Journal of Nursing Studies, vol. 141, p. 104555, 2025.

[3] T. Chu, "The moderating effect of person–job fit on the relationship between compassion fatigue and organizational citizenship behavior among nurses," Journal of Nursing Management, vol. 28, no. 3, pp. 521–529, 2020.

[4] F. M. M. Badran and D. A. Khalaaf, "Work–life balance and its relation to person–job fit among staff nurses," Evidence-Based Nursing Research, vol. 4, no. 3, pp. 86–93, 2022.

[5] I. N. Widayana, I. W. Suardana, and D. A. Mulyani, "Work–life balance and psychosocial well-being among Indonesian nurses: A systematic scoping review," Nurse Media Journal of Nursing, vol. 15, no. 1, pp. 12–28, 2025.

[6] A. Bandura, Self-Efficacy: The Exercise of Control. New York, NY, USA: W. H. Freeman, 1997.

[7] D. R. Compeau and C. A. Higgins, "Computer self-efficacy: Development of a measure and initial test," MIS Quarterly, 1995.

[8] I.-S. Chen, "Computer self-efficacy and learning performance," Computers in Human Behavior, 2017.

[9] Y. Xiao et al., "Person–environment fit and medical professionals' job satisfaction, turnover intention, and professional efficacy: A cross-sectional study in Shanghai," PLOS ONE, vol. 16, no. 4, pp. 1–14, 2021.

[10] M. Engin and F. Gürses, "Adoption of hospital information systems in public hospitals in Turkey: An analysis with the unified theory of acceptance and use of technology model," International Journal of Innovation and Technology Management, vol. 16, no. 6, Art. no. 1950043, 2019.

[11] T. Fukuzaki, N. Iwata, S. Ooba, S. Takeda, and M. Inoue, "The effect of nurses' work–life balance on work engagement: The moderating effect of affective commitment," Yonago Acta Medica, vol. 65, no. 3, pp. 230–239, 2022.

[12] J. Luo et al., "Role of perceived ease of use, usefulness, and financial strength on the adoption of health information systems: The moderating role of hospital size," Humanities and Social Sciences Communications, vol. 11, no. 1, p. 516, 2024.

[13] A. Rhayha, R. Hassan, and N. Mahmoud, “Health workers’ performance in the digital transformation era: The role of technology adoption and system usability,” Journal of Health Management, vol. 27, no. 2, pp. 145–160, 2025.

[14] F. Zheng, L. Xu, Y. Chen, and S. Zhang, “Determinants of clinicians’ utilization of hospital information systems for infection prevention and control: An extended DeLone and McLean model,” JMIR Medical Informatics, vol. 11, no. 5, e41562, 2023.

[15] A. Winter et al., Health Information Systems: Technological and Management Perspectives, 3rd ed. Cham, Switzerland: Springer, 2023.

[16] M. Alghamdi and G. D. Sideridis, “Corrected goodness-of-fit index in latent variable modeling using non-parametric bootstrapping,” Frontiers in Psychology, vol. 16, pp. 1453–1462, 2025.

[17] J. Smeds, A. Johansson, and M. Lundberg, “Professional competence, job fit, and digital self-efficacy among Nordic nurses,” Scandinavian Journal of Caring Sciences, vol. 37, no. 2, pp. 456–468, 2023.

[18] X. Chen, Y. Liu, and J. Zhang, “The impact of person–job fit on nurses’ job satisfaction and performance: The mediating role of organizational commitment,” Journal of Nursing Scholarship, vol. 55, no. 2, pp. 225–234, 2023.

[19] J. R. Edwards and D. M. Cable, “The value of value congruence,” Journal of Applied Psychology, vol. 94, no. 3, pp. 654–677, 2009.

[20] N. Hashim, R. Abdullah, and M. Rahman, “Work–life balance, job satisfaction, and performance among Malaysian nurses in the digital era,” Malaysian Journal of Nursing, vol. 15, no. 1, pp. 45–55, 2023.

[21] A. A. Akanni and C. A. Oduaran, “Person–job fit and work–life balance of female nurses with cultural competence as a mediator: Evidence from Nigeria,” Frontiers of Nursing, vol. 9, no. 1, pp. 81–86, 2022.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1063-1073. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3221

RESEARCH ARTICLE OPEN ACCESS

OPTIMAL ENERGY-EFFICIENT INTEGRATION OF WIND DISTRIBUTED GENERATORS IN A HYBRID AC-DC NETWORK USING A TUNICATE SWARM ALGORITHM

Mohamed Zellagui[1]*, Samir Settoul[2] and Claude Ziad El-Bayeh[3]

[1]Department of Electrical Engineering, University of Batna 2, 05078, Batna, Algeria.
[2]Department of Electrical Engineering, University of Annaba, 23000, Annaba, Algeria.
[3]Department of Electrical Engineering, University of Doha for Science and Technology, 24449, Doha, Qatar.

[1]http://orcid.org/0000-0003-2558-2273, [2]http://orcid.org/0000-0001-7096-7960, [3]http://orcid.org/0000-0002-8268-8878

Email: *m.zellagui@univ-batna2.dz, samir.settoul@univ-annaba.dz, c.bayeh@hotmail.com

ARTICLE INFO

Article History
Received: January 2, 2025
Revised: January 20, 2026
Accepted: January 30, 2026
Published: February 28, 2026

Keywords:
Wind distributed generator,
Hybrid AC/DC system,
Electrical distribution network,
Active power loss,
Tunicate swarm algorithm.

ABSTRACT

Due to improvements in wind energy technology, electrical distribution networks are becoming more capable of integrating large amounts of variable renewable generation. The wind distributed generators (WDGs) affect the medium-voltage (MV) electrical distribution network (EDN) by introducing variability, reverse power flows, power-quality issues, and higher fault levels. This paper applies a brand-new, reliable, and efficient algorithm, the tunicate swarm algorithm (TSA), to optimally link varying numbers of WDG units and allocate multiple WDG units in EDN. The main objective of this research is to propose an applied TSA algorithm for optimal energy-efficient integration of multiple WDG units in a hybrid medium-voltage AC-DC network and to test it in a modified IEEE 33-bus network based on the active power loss minimization. The proposed approach granted improved solution quality with enhanced convergence speed and adherence to EDN constraints. The applied TSA addresses the sensitivity to parameter settings like population size and iterations. The findings verify that, out of all the scenarios taken into consideration, three WDGs offer the best performance in terms of active loss minimization and improve the bus voltage profile.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Because of their many benefits, AC systems have been used in electrical distribution networks (EDNs) for over a century and are now widely used worldwide. However, clean, green, and ecologically friendly distributed energy resources (DERs), electric vehicles, energy storage systems, and diverse DC loads are swiftly revolutionizing the current world [1]. Many advancements in smart grid technology were brought about by the introduction of hybrid AC/DC distribution networks. To connect to current AC distribution systems, these DC-based smart energy technologies require converters, which are costly and incur power losses [2]. In order to directly connect AC and DC generation units and loads to the power system, the idea of medium-voltage (MV) electric distribution networks has recently been proposed. Higher transfer capacity, lower power losses, greater hosting capacity, voltage management, and an improved voltage profile are just a few other benefits of the hybrid AC/DC in EDN [3],[4]. Problems like increased power losses and voltage instability are brought about by the growing integration of wind energy into EDNs. The necessity for strong optimization techniques to guarantee effective and dependable wind distributed generators (WDG) allocation is the driving force behind this project. Electrical distribution grid-connected wind energy systems face persistent challenges in maintaining voltage stability, mitigating harmonics, and ensuring reliable power quality under fluctuating wind conditions and fault events. The intermittent nature of wind introduces significant variations in power flow, which, if not effectively managed, can lead to instability and degrade system performance. Power quality is a crucial aspect of modern electrical distribution grids, ensuring efficient, stable, and reliable power delivery while minimizing losses and disturbances.

With the growing integration of wind energy into grid-connected systems, Power quality issues have become more pronounced due to the inherent intermittency and variability of WDG. In recent years, many scholars have proposed numerous optimizations approaches to formulate the WDG planning problem in distribution systems. Numerous optimization approaches have been developed to address the efficient design of WDG units in EDN. The authors of [5] applied the equilibrium optimizer algorithm, the gorilla troops optimizer algorithm in [6], the lévy coyote optimization algorithm in [7], and used the manta-ray foraging optimization algorithm in [8] and modified the equilibrium algorithm in [9]. This paper introduces a novel bio-inspired optimization technique called the Tunicate Swarm Algorithm (TSA) [10]. This algorithm draws inspiration from the jet propulsion and collective foraging behaviors exhibited by tunicates, aiming to enhance navigation and foraging processes through innovative computational strategies. The TSA algorithm presents a straightforward framework, unique stabilizing characteristics, and remarkable efficiency [11]. The TSA algorithm surpasses competing methods in optimal solutions and highly effective for real-world optimization situations, and especially for power system problems: optimize the power system stabilizer parameters to prevent oscillations at low frequencies [12], used to address the dynamic economic emission dispatch issues [13],[14], the PID controller optimization for autonomous generation management in a hybrid power system [15], Optimal cost of a power system based on the design of a hybrid renewable system [16], and optimal renewable-based DG in electrical networks with electric load models [17]. Adaptive optimal control model-based optimal energy management strategy for a multi-microgrid at the tertiary level [18], optimizing battery energy storage systems and vehicle-to-grid operations in PV-powered EV charging systems [19], and optimal solutions searched for wind speed deterministic forecasting [20]. Recently, researchers applied the TSA algorithm for energy hub adjusting efficiency based on real-time load variations for optimal allocation [21]. The main objective of this research is to propose an applied TSA algorithm for optimal energy-efficient integration of multiple wind distributed generators in a hybrid medium-voltage AC-DC electric distribution network. The proposed approach ensures improved solution quality with enhanced convergence speed and adherence to EDN constraints. This, in turn, ultimately improves the stability, efficiency, and reliability of EDN. Through the balancing strategy between exploration and exploitation, TSA addresses the sensitivity to parameter settings like population size and iterations.

II. WIND DISTRIBUTED GENERATOR ALLOCATION PROBLEM

In EDN, computational solutions are crucial because of the intricate nature of real-time issues. This involves formulating the problem through mathematical equations, known as mathematical models, which are predicated on decision variables. These models precisely illustrate the system's response to variations in independent variables, guaranteeing relevance and accuracy in practical applications. Figure 1 represents the single-line diagram with DG placement in EDN.

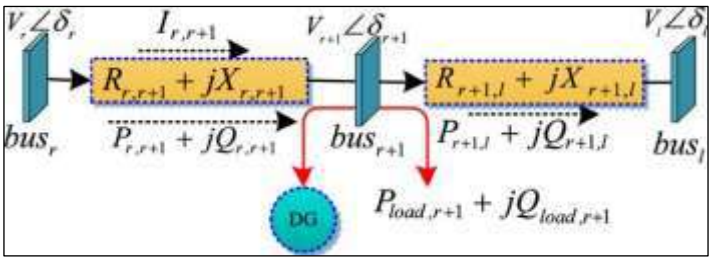


Figure 1: The single-line diagram with DG placement.
Source: [18-20].

II.1 OBJECTIVE FUNCTION

To achieve optimal integration of WDG units within the hybrid AC-DC EDN, the problem is formulated using a single objective function (OF), which aims to minimize the total active power losses (P_{Loss}), as expressed by the following equation:

$$OF = \sum_{i=1}^{N_{Bus}} \sum_{j=2}^{N_{Bus}} P_{Loss}(i,j) \tag{1}$$

The active power loss is calculated using the expression given in (2), as reported in [6],[22]:

$$P_{Loss}(i,j) = \alpha_{ij}\left(P_i P_j + Q_i Q_j\right) + \beta_{ij}\left(Q_i P_j + P_i Q_j\right) \tag{2}$$

And,

$$\alpha_{ij} = \frac{R_{ij}}{V_i V_j}\cos\left(\delta_i - \delta_j\right) \tag{3}$$

$$\beta_{ij} = \frac{R_{ij}}{V_i V_j}\sin\left(\delta_i - \delta_j\right) \tag{4}$$

Where R_{ij} denotes the resistance of the line connecting buses i and j. V_i and V_j are the voltage magnitudes. δ_i and δ_j represent the corresponding voltage angles. P_i, P_j and Q_i, Q_j refer to the power injections at buses i and j, respectively.

II.2 EQUALITY CONSTRAINTS

These constraints enforce active and reactive power balance in the system, where total generation (including WDG) equals load demand plus network losses, as defined in (5) and (6) [23],[34]:

$$P_G + P_{WDG} = P_{Load} + P_{Loss} \quad (5)$$

$$Q_G + Q_{WDG} = Q_{Load} + Q_{Loss} \quad (6)$$

Where P_G and Q_G denote the active and reactive power outputs of the generator, respectively. P_{WDG} and Q_{WDG} represent the total power supplied by the WDG unit. P_{Load} and Q_{Load} correspond to the total active and reactive load demands, while P_{Loss} and Q_{Loss} indicate the total active and reactive power losses, respectively.

II.3 DISTRIBUTION LINE CONSTRAINTS

These constraints ensure acceptable AC-DC distribution line operating conditions [23, 24]:

$$V_{min} \leq V_i \leq V_{max} \quad (7)$$

$$|1 - V_j| \leq \Delta V_{max} \quad (8)$$

$$|S_{ij}| \leq S_{max} \quad (9)$$

Where V_{max}, and V_{min} are the limit voltages; ΔV_{max} is the maximum drop of voltage. V_1 is the sub-station voltage equal to 1.0 p.u. S_{max} is the maximum apparent power in line.

II.4 WDG CONSTRAINTS

These constraints specify the operational and placement limits associated with WDG units, as defined in [23],[24]:

$$P_{WDG}^{min} \leq P_{WDG} \leq P_{WDG}^{max} \quad (9)$$

$$Q_{WDG}^{min} \leq Q_{WDG} \leq Q_{WDG}^{max} \quad (10)$$

$$2 \leq WDG_{position} \leq N_{Bus} \quad (11)$$

$$N_{WDG} \leq N_{WDG.max} \quad (12)$$

$$n_{WDG}/Location \leq 1 \quad (13)$$

Where $WDG_{Position}$, N_{WDG}, and $N_{WDG.max}$ are the WDG position, number, and maximum units/location at the bus.

III. OVERVIEW OF APPLIED TUNICATE SWARM ALGORITHM

The TSA is predicated on tunicates' capacity to locate food sources in the ocean. Here, tunicates use two behaviours, such as swarm intelligence and jet propulsion actions, to locate food. Three requirements are considered for the implementation of jet propulsion behaviour: the search candidates must avoid conflicts with one another, approach the position of the best candidate, and exist close to the best candidate [10, 11]. The vector $\vec{\beta}$ is used to avoid conflicts among the tunicates, and the position of the new search candidate is calculated using this vector [25, 26]:

$$\vec{\beta} = \frac{\vec{g}}{\vec{m}} \quad (14)$$

$$\vec{g} = c_2 + c_3 - \vec{f} \quad (15)$$

$$\vec{f} = 2.c_1 \quad (16)$$

Here, $\vec{f}$ indicates the water flow in the sea, $\vec{m}$ indicates the force existing within search candidates, $\vec{g}$ denotes the force of gravityc_1, c_2, c_3 are the parameters distributed randomly between 0 and 1.

$$\vec{m} = [p_{min} + c_1.p_{max} - p_{max}] \quad (17)$$

Here, p_{min}, p_{max} values are given by 1 and 4, respectively and indicate the interaction speeds of search candidates. The tunicates move to the best neighbor search candidate after avoiding conflicts, and this is given by the next equation [24]:

$$\vec{d} = |\vec{f_s} - r.\vec{p}\,(t)| \quad (18)$$

Where, $\vec{p}$ indicates the location of the search candidate for the current location, t indicates the value of the current iteration, $\overrightarrow{d}$ represents the distance between the tunicate and the food location, $\overrightarrow{f_s}$ denotes the position of food, and r ranges between 0 and 1. The location of the search candidate has to be adjacent to the best search candidate, which is given by,

$$\vec{p}\,(t) = \begin{cases} \overrightarrow{f_s} + \vec{\beta}.\overrightarrow{d}, r \geq 0.5 \\ \overrightarrow{f_s} - \vec{\beta}.\overrightarrow{d}, r < 0.5 \end{cases} \tag{19}$$

Here, in accordance with $\overrightarrow{f_s}$, the updating of the search candidate's position is given by $\vec{p}\,(t)$. The position updating is achieved by the swarm intelligence behaviour of the search candidates and is,

$$\vec{p}\,(t+1) = \frac{\vec{p}\,(t)+\vec{p}\,(t+1)}{2+c_1} \tag{20}$$

The solution of the first two best candidates is given by $\vec{p}\,(t)$ and $\vec{p}\,(t+1)$. Considering the position of the best candidate, the other candidate's solution can be updated by preserving these solutions. Figure 2 shows how search agents can update their positions according to the position of $\vec{p}\,(t)$. The final position would be in a random place, within a cylindrical or cone-shaped, which is defined by the position of tunicate [25].

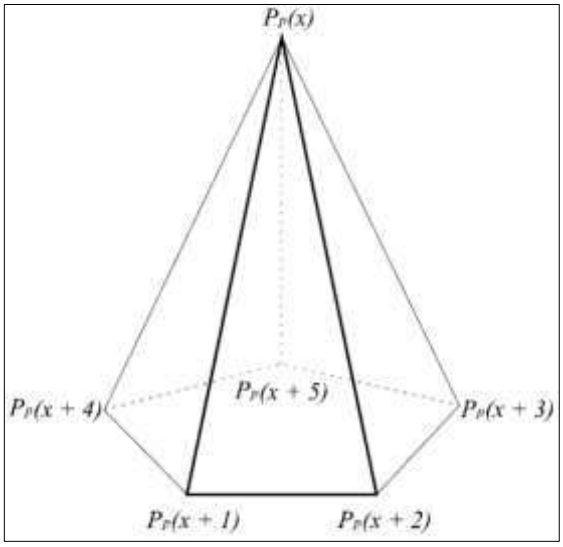


Figure 2: Presentation of the 3D position vectors of the tunicate.
Source: [25].

The parameters, positions and velocities are continuously updated till the convergence criteria are met. At each iteration, update the global best solution based on the fitness evaluations [26]. Consequently, preserve the solutions of the first two best candidates and use them to update other candidates' solutions while maintaining swarm intelligence behavior.

Table 1: Parameters of Applied TSA.

Parameters	Values
c_1	0.5
c_2	0.3
c_3	0.7
p_{min}	1
p_{max}	4
Population size	50
Maximum iterations	200

Source: Authors, (2026).

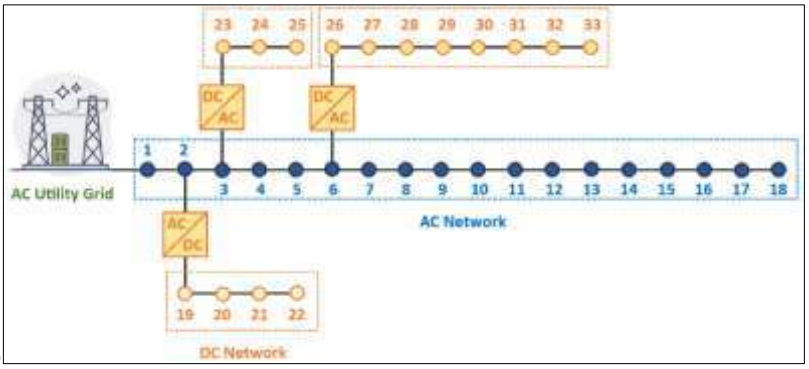
IV. RESULTS AND DISCUSSIONS

In this section, an applied optimization TSA technique for the optimum location of multiple WDG units is utilized to lessen power loss. The proposed controller is executed on MATLAB software. The performance of the proposed system is analyzed with a modified IEEE 33-bus system represented in Figure 3, which was adopted for the validation of the proposed model. The illustration demonstrates the one-line representation of the modified IEEE 33 bus distribution system, designed as a hybrid AC-DC network. It emphasizes the incorporation of DC sub-networks into a traditional radial AC feeder using power electronic interfaces. The AC utility grid provides power to the system via the substation at Bus 1, preserving the radial configuration of the 33-bus feeder. Buses 1 through 18 constitute the AC distribution network, illustrated in blue.

In addition to providing traditional AC loads, this portion acts as the framework for linking DC subsystems. Practical MV distribution systems are reflected in the radial structure. The AC system is linked to two DC networks (illustrated in orange). DC loads and DC-based distributed resources, including wind distributed generation, energy storage systems, DC microgrids, and DC consumption, are represented by these DC sections. This one-line diagram serves as the test system architecture for evaluating voltage profiles, power flow, and the impact of WDG penetration in a hybrid network. It provides the structural basis for the voltage performance results observed in the different case studies (cases 1 to 4). The four configurations for case studies in this paper:

- Case 1: Hybrid EDN without WDG (basic case),
- Case 2: Hybrid EDN with one WDG,
- Case 3: Hybrid EDN with two WDGs,
- Case 4: Hybrid EDN with three WDGs.

Figure 3: One-line diagram of the tested modified IEEE 33-bus hybrid AC-DC network.



Source: Authors, (2026).

Figure 4 represents the bus voltage of a hybrid AC-DC network without WDG units (case 1).

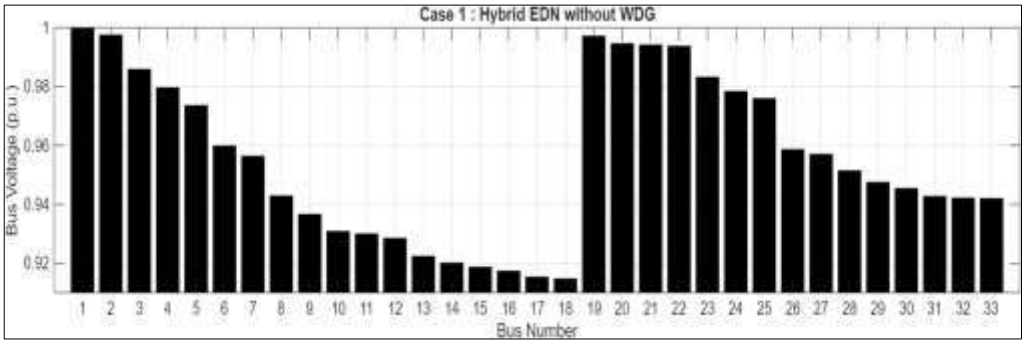


Figure 4: The bus voltage of a hybrid AC-DC network without WDG units (Case 1).
Source: Authors, (2026).

As observed in Figure 4, because the feeder is radial, the bus voltage magnitude gradually drops from the substation (Bus 1) to the end buses. Some buses suffer from severe undervoltage, with values falling below 0.95 p.u., especially in the downstream AC network (around buses 9–18). Voltages in the DC network (buses 19–33) likewise progressively drop, suggesting greater vulnerability to loading and inadequate voltage support. From the standpoint of power quality, this instance is crucial since the voltage profile. Figure 5 represents the branch active power loss of a hybrid AC-DC network without WDG units (case 1).

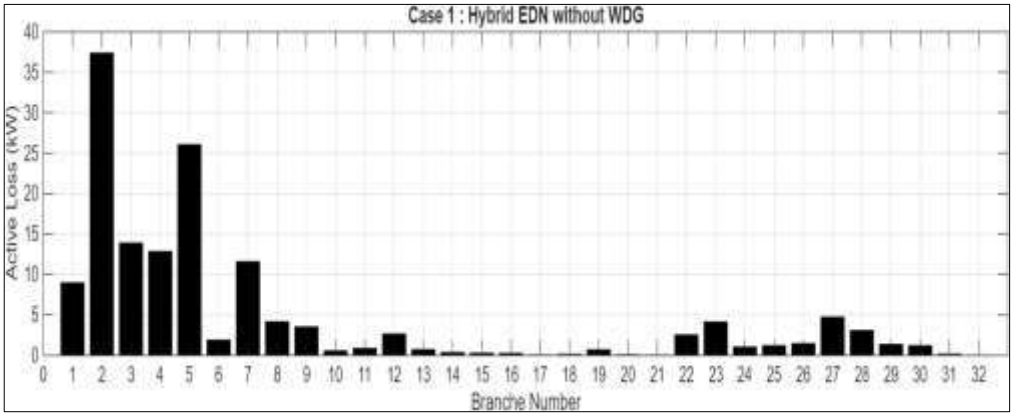


Figure 5: The branch active power loss of a hybrid AC-DC network without WDG units (Case 1).
Source: Authors, (2026).

As demonstrated by Figure 5, the upstream branches, particularly those near the substation (e.g., branches 2–6), have a significant concentration of active power losses. This is because there is no local production to balance the electricity demand, and the entire downstream load is supplied by a high current flow. Even while losses in downstream branches are very minor, they nevertheless add to the overall inefficiency of the system. As a result, the overall active power loss is comparatively large. The branch reactive power loss of a hybrid AC-DC network without WDG units (case 1) is shown in Figure 6.

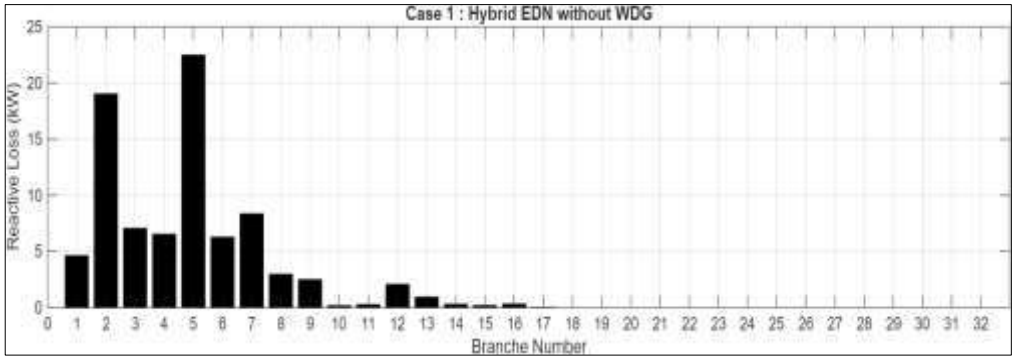


Figure 6: The branch reactive power loss of a hybrid AC-DC network without WDG units (Case 1).
Source: Authors, (2026).

As shown in Figure 6, with peak losses in the first feeder branches, reactive power losses exhibit a similar pattern to active losses. Large reactive power demands from loads are indicated by high reactive losses, which worsen voltage drops along the feeder. The system has very little reactive support, which exacerbates voltage deterioration. The convergence characteristics of the proposed TSA algorithm after integrating one, two, and three WDG units in EDN are presented in Figure 7.

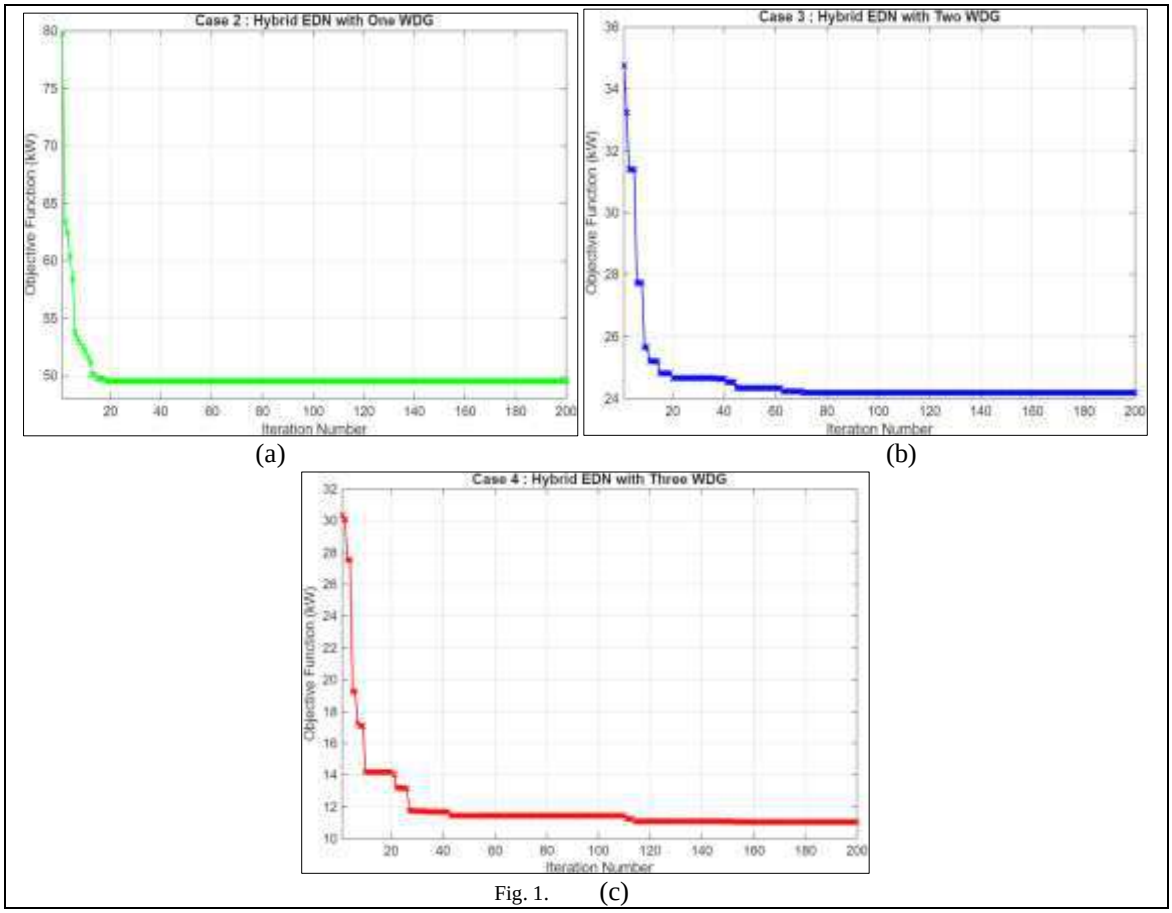


Fig. 1.

Figure 7: The convergence curves of the TSA approach for all case studies:
a). Case 1: One WDG, b). Case 2: Two WDGs, c). Case 3: Three WDGs.
Source: Authors, (2026).

The proposed approach ensures improved solution quality with enhanced convergence speed and adherence to EDN constraints. This, in turn, ultimately improves the stability, efficiency, and reliability of the test system. At each iteration, update the best global solution based on the objective function. Through the balancing strategy between exploration and exploitation, TSA addresses the sensitivity to parameter settings like population size and iterations for the three case studies. Tables 2, 3, and 4 summarize the optimal parameters after installing one, two, and three WDG units and technical parameters for the hybrid AC-DC test system.

Table 2: Optimal Planning of One WDG unit (Case 2).

Parameters		Values
WDG Unit	***Bus Location***	6
	P_{WDG} **(MW)**	2.5234
	Q_{WDG} **(MVar)**	0.6507
Hybrid AC-DC Network	ΣP_{Loss} **(MW)**	49.5134
	ΔP_{Loss} **(%)**	66.5456
	ΣQ_{Loss} **(MVar)**	24.9955
	ΔQ_{Loss} **(%)**	70,3752
	V_{min} **(p.u.)**	0.9578
	V_{max} **(p.u.)**	1.0009
	Σ ***VD*** **(p.u.)**	0.4876

Source: Authors, (2026).

Table 3: Optimal Planning of Two WDG Units (Case 3).

Parameters		Values
WDG Units	***Bus Location***	13 – 30
	P_{WDG} **(MW)**	0.8424 – 1.1210
	Q_{WDG} **(MVar)**	0.3734 – 0.1058
Hybrid AC-DC Network	ΣP_{Loss} **(MW)**	24.1855
	ΔP_{Loss} **(%)**	83.6587
	ΣQ_{Loss} **(MVar)**	7.6940
	ΔQ_{Loss} **(%)**	90.8810
	V_{min} **(p.u.)**	0.9847
	V_{max} **(p.u.)**	1.0001
	Σ ***VD*** **(p.u.)**	0.1778

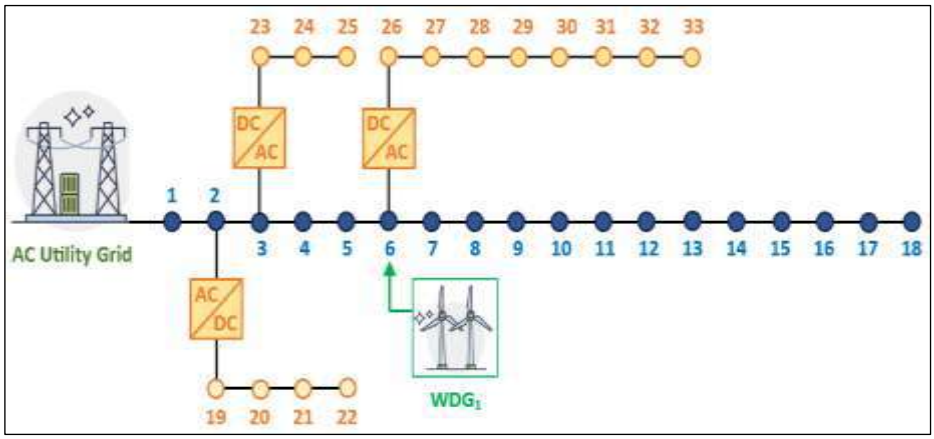
Source: Authors, (2026).

Table 4: Optimal Planning of Three WDG Units (Case 4).

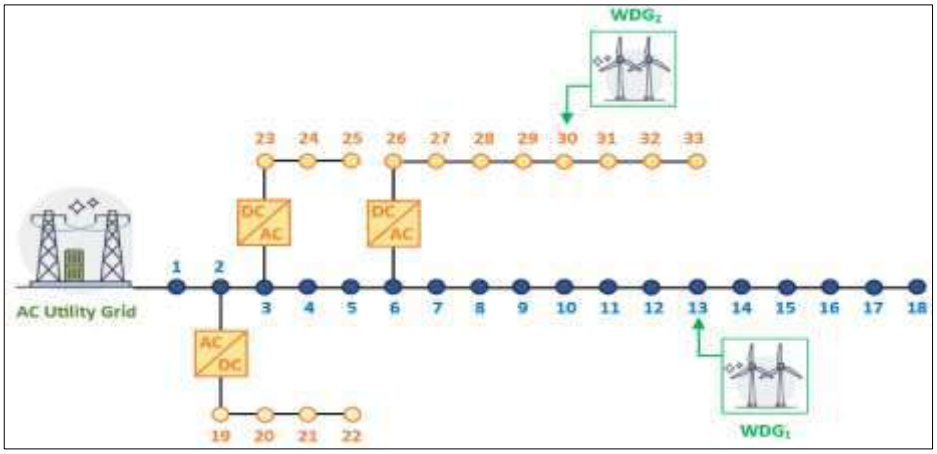
Parameters		Values
WDG Units	***Bus Location***	14 – 24 – 30
	P_{WDG} **(MW)**	0.8059 – 1.1683 – 0.9941
	Q_{WDG} **(MVar)**	0.3671 – 0.1955 – 0.1118
Hybrid AC-DC Network	ΣP_{Loss} **(MW)**	11.0467
	ΔP_{Loss} **(%)**	92.5361
	ΣQ_{Loss} **(MVar)**	4.6155
	ΔQ_{Loss} **(%)**	94.5297
	V_{min} **(p.u.)**	0.9929
	V_{max} **(p.u.)**	1.0034
	Σ ***VD*** **(p.u.)**	0.0803

Source: Authors, (2026).

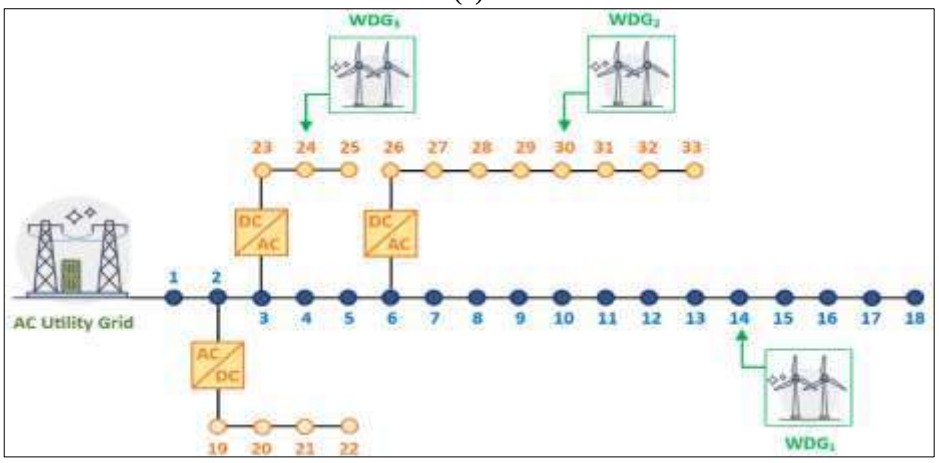
Tables 1, 2, and 3 demonstrate the benefits of installing WDG units, particularly in terms of reduced power losses and voltage deviation. Furthermore, compared to other situations, the integration of the three WDG produces a good performance. The configuration for the hybrid AC-DC systems with the three case studies, with location one, two, and three WDG units, is presented in Figure 7.



(a)



(b)



(c)

Figure 8: Optimal configuration of WDG integration for various case studies:
a). Case 1: One WDG, b). Case 2: Two WDGs, c). Case 3: Three WDGs.
Source: Authors, (2026).

Figure 9 compares bus voltage magnitudes (p.u.) across the hybrid AC-DC network for different case studies with increasing numbers of WDG units.

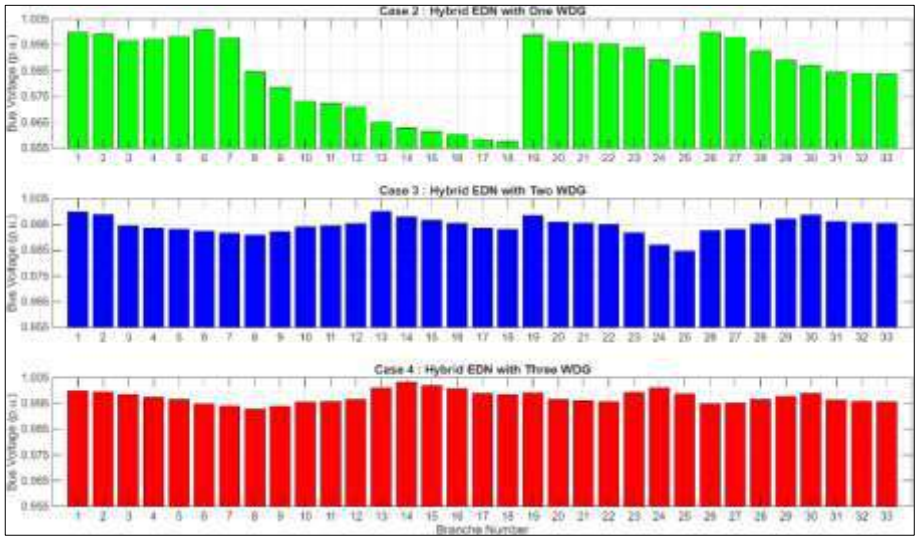


Figure 9: The Bus voltage of a hybrid AC-DC network with WDG units for various case studies.
Source: Authors, (2026).

After analyzing Figure 9. In case 2, the hybrid EDN with one WDG (green color): The voltage profile shows larger variations along the feeder. Several downstream buses (roughly buses 9–18) experience noticeable voltage drops, approaching the lower acceptable

limit (~0.955–0.965 p.u.). This indicates that a single WDG provides limited voltage support, especially in buses far from the source or in load-intensive sections. Voltage regulation is weaker and more location-dependent.

For Case 3, the hybrid EDN with two WDG units (blue color), the voltage profile becomes significantly flatter than in Case 2. Most bus voltages lie within a narrower band (≈0.985–1.0 p.u.). Localized voltage dips are reduced, showing that additional WDG improves reactive power support and reduces line voltage drops. Overall voltage regulation and uniformity are improved. Finally, in Case 4, the hybrid EDN with three WDG units (red color): This case exhibits the best voltage profile among all scenarios. Bus voltages are very close to the nominal value (≈1.0 p.u.) across the entire network. Voltage deviations are minimal, and no critical undervoltage buses are observed. The results highlight that higher WDG penetration enhances voltage stability and reduces losses in the hybrid AC–DC network. In conclusion, the figure clearly demonstrates the positive impact of coordinated WDG integration on bus voltage performance in a hybrid AC-DC distribution system. The branch active power loss (kW) over the hybrid AC-DC network for several case studies with an increase in WDG units is compared in Figure 10.

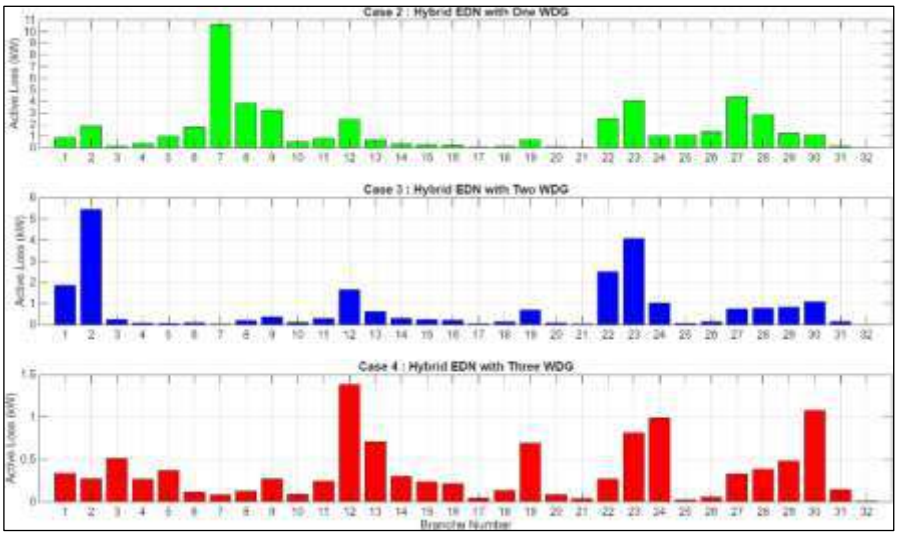


Figure 10: The branch active loss of a hybrid AC-DC network with WDG units for various case studies.
Source: Authors, (2026).

Out of all the possibilities, case 2 shows the greatest branch losses. Some branches have extremely large loss peaks, particularly those near branches 6–8 and 12. Higher power flow from the main grid and increased current in several branches result from the single WDG's restricted local power support. In the hybrid network, one WDG is not enough to considerably reduce branch losses. When comparing Case 3 to Case 2, a significant decrease in branch losses is seen. Peak losses are distributed more uniformly among branches and are substantially lower. In Case 4, the arrangement of three WDGs exhibits the lowest branch losses of any branch. Loss values are uniformly distributed and quite modest. This shows that the hybrid AC-DC network is making the best use of distributed generation. Near-optimal loss minimization and the best-balanced network functioning are achieved by three WDG units. Figure 11 illustrates the branch reactive power losses in the hybrid AC–DC distribution network for different case studies with an increasing number of WDG units.

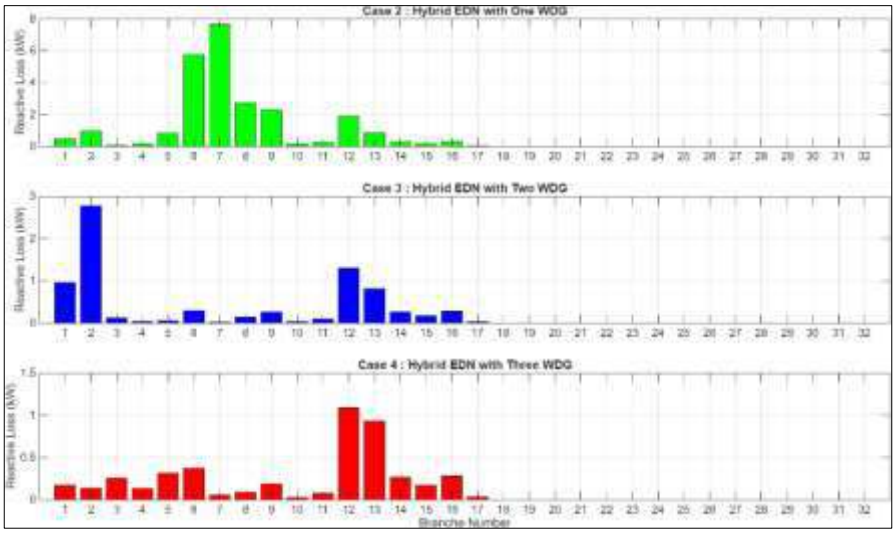
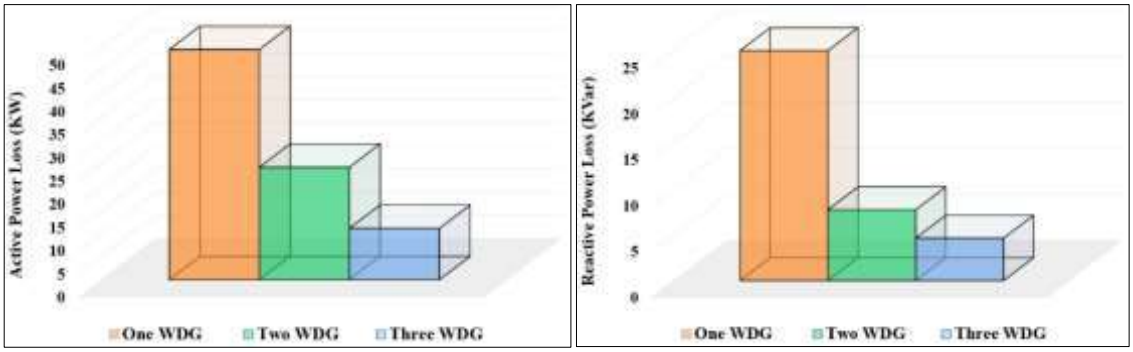


Figure 11: The Branch reactive loss of a hybrid AC-DC network with WDG units for various case studies.
Source: Authors, (2026).

Out of all cases, Case 2 has the largest reactive power losses. There are noticeable loss peaks in several branches, especially in the mid-feeder and upstream parts (around branches 6–8 and 12). The losses are more uniformly distributed, and the peak loss magnitudes

are significantly smaller in Case 3. Better local reactive power adjustment is provided by the extra WDG. However, some branches continue to exhibit moderate losses, indicating partial but incomplete optimization.

Out of all the branches, Case 4 has the lowest reactive power losses. There are no sharp peaks in the loss numbers, which are minor and evenly dispersed. Increased WDG penetration improves both overall reactive current flow reduction and local reactive power support. Figure 12 represents the total active and reactive power losses in a hybrid AC-DC network with increasing WDG penetration.



(a) (b)

Figure 12: The total losses in the hybrid AC-DC network for various WDG units: a). Active power loss, b). Reactive power loss.
Source: Authors, (2026).

Below is a clear technical analysis of Figures 12 (a) and (b) showing the total losses in a hybrid AC–DC network with increasing WDG penetration. The active power loss achieves its lowest value when three WDGs are integrated into EDN, as shown in Figure 12.a, indicating the most successful loss minimization of the situations under study. Increasing the number of WDG devices lowers line currents by improving local power supply and voltage support. As WDG penetration rises, the graph unmistakably demonstrates a monotonic decline in active power losses. More WDG units raise the overall power factor, decrease needless reactive power circulation, and increase local reactive power support. Reactive losses are significantly reduced as a result, which is essential for voltage stability. The findings verify that, out of all the scenarios taken into consideration, three WDGs offer the best performance in terms of loss minimization.

V. CONCLUSIONS

This paper studies the minimization of active power losses by the optimal allocation of multiple WDG units based on applying the TSA algorithm and tested on the modified IEEE 33-bus with a hybrid AC-DC distribution network. In conclusion, the figure clearly demonstrates the positive impact of coordinated WDG integration on bus voltage performance in EDN. The optimal results clearly demonstrate that higher WDG penetration significantly improves power loss performance in the hybrid AC–DC distribution network. The configuration with three WDG units provides the most efficient and stable operation, validating the effectiveness of WDG in reducing reactive losses and enhancing overall network efficiency. These simulation results clearly demonstrate that strategic integration of multiple WDG units in a hybrid AC–DC EDN leads to substantial loss reduction and improved power flow performance, validating the effectiveness of distributed renewable generation in modern EDN. Increasing the number of WDG units leads to a progressive reduction in branch power losses. The hybrid AC–DC network benefits substantially from distributed active and reactive power support, especially when WDGs are optimally integrated.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Mohamed Zellagui, Samir Settoul and Claude Ziad El-Bayeh.
Methodology: Samir Settoul.
Investigation: Mohamed Zellagui and Samir Settoul.
Discussion of results: Mohamed Zellagui and Samir Settoul.
Writing – Original Draft: Mohamed Zellagui and Samir Settoul.
Writing – Review and Editing: Mohamed Zellagui, Samir Settoul and Claude Ziad El-Bayeh.
Resources: Mohamed Zellagui, Samir Settoul and Claude Ziad El-Bayeh.
Supervision: Mohamed Zellagui.
Approval of the final text: Mohamed Zellagui, Samir Settoul and Claude Ziad El-Bayeh.

VII. REFERENCES

[1] S. J. U. Hassan, A. Mehdi, Z. Haider, J. S. Song, A. D. Abraham, G. S. Shin, and C. H. Kim, "Towards medium voltage hybrid AC/DC distribution Systems: architectural topologies, planning and operation", International Journal of Electrical Power & Energy Systems, vol. 159, 110003, 2024. https://doi.org/10.1016/j.ijepes.2024.110003.

[2] S. Charadi, Y. Chaibi, A. Redouane, A. Allouhi, A. El Hasnaoui, and H. Mahmoudi, "Efficiency and energy-loss analysis for hybrid AC/DC distribution systems and microgrids: A review", International Transactions on Electrical Energy Systems, vol. 31, e13203, 2021. https://doi.org/10.1002/2050-7038.13203.

[3] Y. Chen, Q. Lu, K. Zeng, Y. Yang, and P. Xie, "Flexibility-oriented AC/DC hybrid grid optimization using distributionally robust chance-constrained method", Energies, vol. 17, no. 19, 4902, 2024. https://doi.org/10.3390/en17194902.

[4] M. Zellagui, N. Belbachir, A. Lasmari, R. J. J. Molu, and S. Kamel, "Enhancing PV distributed generator planning in medium-voltage DC distribution networks: a multi-design techno-economic analysis with load demand response", IET Generation, Transmission & Distribution, vol. 18, no. 1, pp. 173–189, 2024. https://doi.org/10.1049/gtd2.13080.

[5] M. Zellagui, S. Settoul, and A. Lasmari, “Applied equilibrium optimizer algorithm for optimal planning multiple photovoltaic distributed generators in hybrid AC-DC distribution networks, Brazilian Journal of Technology, vol. 7, no. 4, e76100, 2024. https://doi.org/10.38152/bjtv7n4-042.

[6] N. Belbachir, M. Zellagui, S. Settoul, and A. Benali, “Optimal design of wind energy generation in electricity distribution network based technical-economic parameters using metaheuristic optimization algorithms”, Engineering Review, vol. 43, no. 3, pp. 115–128, 2023. https://doi.org/10.30765/er.2281.

[7] E. Kaymaz, S. Duman, and U. Guvenc, “Optimal power flow solution with stochastic wind power using the Lévy coyote optimization algorithm”, Neural Computing and Applications, vol. 33, no. 12, pp. 6775–6804, 2021. https://doi.org/10.1007/s00521-020-05455-9.

[8] A. K. Khamees, A. Y. Abdelaziz, M. R. Eskaros, H. H. Alhelou, and M. A. Attia, “Stochastic modeling for wind energy and multi-objective optimal power flow by novel meta-heuristic method”, IEEE Access, vol. 9, pp. 158353–158366, 2021. https://doi.org/10.1109/access.2021.3127940.

[9] M. Q. Duong, T. Nguyen, and T. T. Nguyen, “Optimal placement of wind power plants in transmission power networks by applying an effectively proposed metaheuristic algorithm”, Mathematical Problems in Engineering, vol. 2021, e1015367, 2021. https://doi.org/10.1155/2021/1015367.

[10] S. Kaur, L. K. Awasthi, A. Sangal, and G. Dhiman, “A new bio-inspired based metaheuristic paradigm for global optimization”, Engineering Applications of Artificial Intelligence, vol. 90, 103541, 2020. https://doi.org/10.1016/j.engappai.2020.103541.

[11] H. Jumakhan, S. Abouelnour, A. Al-Redhaei, S. N. Makhadmeh, and M. A.Al-Betar, “Recent versions and applications of tunicate swarm algorithm”, Archives of Computational Methods in Engineering, vol. 32, pp. 4857–4886, 2025. https://doi.org/10.1007/s11831-025-10287-8.

[12] M. R. Islam, M. S. Azam, M. S. Hossen, M. S. Islam, M. Y. Worku, M. S. Shahriar, and M. Shafiullah, “Power system stability enhancement through optimal PSS design”, E-Prime - Advances in Electrical Engineering, Electronics and Energy, vol. 9, 100735, 2024. https://doi.org/10.1016/j.prime.2024.100735.

[13] L. L. Li, Z. F. Liu, M. L. Tseng, S. J. Zheng, and M. K. Lim, “Improved tunicate swarm algorithm: solving the dynamic economic emission dispatch problems”, Applied Soft Computing, vol. 108, 107504, 2021. https://doi.org/10.1016/j.asoc.2021.107504

[14] R. M. Rizk-Allah, E. A. Hagag, and A. A. El-Fergany, “Chaos-enhanced multi-objective tunicate swarm algorithm for economic-emission load dispatch problem”, Soft Computing, vol. 27, pp. 5721–5739, 2023. https://doi.org/10.1007/s00500-022-07794-2.

[15] A.S. Memon, J. Laghari, M.A. Bhayo, S. Khokhar, S. Chandio, and M.S. Memon, “Tunicate swarm algorithm based optimized PID controller for automatic generation control of two area hybrid power system”, Journal of Intelligent & Fuzzy Systems, vol. 45, no. 2, pp. 2565–2578, 2023. https://doi.org/10.3233/JIFS-223227.

[16] R. Krishnakumar, and C. Ravichandran, “Reliability and cost minimization of renewable power system with tunicate swarm optimization approach based on the design of PV/Wind/FC system”, Renewable Energy Focus, vol. 42, pp. 266–276, 2022. https://doi.org/10.1016/j.ref.2022.07.003.

[17] H. S. Madupu, P. R. Chinda, and S. Kotni, “A novel tunicate swarm algorithm for optimal integration of renewable distribution generation in electrical distribution networks considering extreme load growth”, Journal of Electrical Engineering & Technology, vol. 18, pp. 2709–2722, 2023. https://doi.org/10.1007/s42835-023-01388-0.

[18] R. Rashidi, A. Hatami, M. Moradi, and X. Liang, “Optimal multi-microgrids energy management through information gap decision theory and tunicate swarm algorithm”, IEEE Access, vol. 12, pp. 114795–114808, 2024. https://doi.org/10.1109/access.2024.3443471.

[19] P. Hemachandu, S. Premalatha, K. P. Arunachalam, and P. V. H. Prasad, “Energy management in photovoltaic-based electric vehicle charging systems with battery storage and vehicle-to-grid support”, Energy Storage, vol. 7, no. 5, e70236, 2025. https://doi.org/10.1002/est2.70236.

[20] J. Wang, S. Wang, and Z. Li, “Wind speed deterministic forecasting and probabilistic interval forecasting approach based on deep learning, modified tunicate swarm algorithm, and quantile regression”, Renewable Energy, vol. 179, pp. 1246–1261, 2021. https://doi.org/10.1016/j.renene.2021.07.113.

[21] Q. Su, Z. Gao, and Z. Zhong, “Dynamic energy hub-based planning and operation for optimal integration of variable energy conversion efficiency in community-level energy systems”, Energy, vol. 332, 137143, 2025. https://doi.org/10.1016/j.energy.2025.137143

[22] H. A. El-Sattar, S. Kamel, and M. A. Elseify, “A modified white shark optimizer for optimizing photovoltaic, wind turbines, biomass, and hydrogen storage hybrid systems”, Journal of Energy Storage vol. 113, 115655, 2025. https://doi.org/10.1016/j.est.2025.115655.

[23] S. Settoul, M. Zellagui, and R. Chenni, “A new optimization algorithm for optimal wind turbine location problem in Constantine city electric distribution network based active power loss reduction”, Journal of Optimization in Industrial Engineering, vol. 14, no. 2, pp. 13 –22, 2021. https://doi.org/10.22094/joie.2020.1892184.1725.

[24] N. Belbachir, M. Zellagui, and S. Kamel, “Multi-dimensional energy management based on an optimal allocation of hybrid wind turbine distributed generation and battery energy storage system in a flexible interconnected distribution network considering seasonal uncertainties”, Engineering Proceedings, vol. 56, no. 1, 246, 2023. https://doi.org/10.3390/ASEC2023-16292.

[25] O. Akdağ, “A modified tunicate swarm algorithm for engineering optimization problems”, Arabian Journal for Science and Engineering, vol. 48, no. 11, pp. 14745–14771, 2023. https://doi.org/10.1007/s13369-023-07803-y.

[26] R. Zheng, A.G. Hussien, A. Bouaouda, R. Zhong, G. Hu, “A comprehensive review of the tunicate swarm algorithm: variations, applications, and results”, Archives of Computational Methods in Engineering, vol. 32, pp. 2917-2986, 2025. https://doi.org/10.1007/s11831-025-10228-5.

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1074-1083. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3235

RESEARCH ARTICLE OPEN ACCESS

OF AHERO RICE SURVIVAL ANALYSIS WITH CEMPETING RISKS: INSIGHTS FROM THE FINE AND GRAY MODEL AND COVARIATE ADJUSTMENTS

Sumithra B[1], Anu Swedha Ananthan[2], Zechariah Jebakumar Arulanantham[3]

[1*]Assistant Professor, Department of Mathematics, Faculty of Science & Humanities, SRM Institute of Science and Technology, Ramapuram, Chennai – 600089, Tamil Nadu, India.

[2] Associate Professor, Department of Microbiology, Justice Basheer Ahmed Sayeed College for Women, Teynampet, Chennai – 600018, Tamil Nadu, India.

[3] Lecturer for Biostatistics, Vice Deanship of Postgraduate Studies and scientific research, Prince Sultan Military college of Health science, AD Dammam 343137761, Kingdom of Saudi Arabia.

[1]http://orcid.org/0000-0001-9384-9318, [2]http://orcid.org/0000-0002-4965-3278, [3]http://orcid.org/0000-0002-5978-2436

Email: *sumithrb@srmist.edu.in, anuswedha.a@jbascollege.edu.in, zacbiostat@gmail.com

ARTICLE INFO

Article History

Received: January 4, 2025

Revised: January 20, 2026

Accepted: January 30, 2026

Published: February 28, 2026

Keywords:

Competing Risks

Fine and Gray Model

Survival Analysis

Cumulative incidence Function

Kaplan-Meier Survival

ABSTRACT

Competing risk survival analysis is deemed necessary when manifold events, such as recurrence and death, are present to prevent the incidence of the event of interest. The application of the Fine and Gray model for cumulative incidence function (CIF) estimation and competing risk regression in a simulated dataset is explored in this study. Simulated time-to-event data and covariates are used to estimate the Fine and Gray model both with and without covariates, with comparisons being made against the Kaplan-Meier (KM) method, which does not account for competing risks. It is demonstrated that the Fine and Gray method provides a more accurate representation of event-specific incidence in the presence of competing risks. Competing risks being ignored, as in the Kaplan-Meier approach, can lead to an underestimation of the event of interest. Additionally, the inclusion of covariates like age and tumor size into the Fine and Gray model is shown to significantly impact the incidence of recurrence and competing events. This comparative analysis is provided.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

I.1 COMPETING RISKS

In survival analysis, the primary focus is typically placed on the time awaiting an event to occur, such as death, failure of a system, or relapse in disease. However, competing risks are encountered when multiple types of events can be practiced by individuals in a study, and the occurrence of one event may avert the occurrence of others. For example, in a cancer study, a patient may die from cancer (the primary event of interest) or from another cause like a heart attack (a competing event). In the presence of competing risks, traditional survival analysis methods must be adjusted to account for the fact that competing events can censor or modify the probability of the primary event being observed.

I.2 KM ESTIMATOR

The KM estimator is a non-parametric statistic that is used to estimate the survival function from life span data. An estimate of the probability that an individual will survive further than a certain time point is offered by it. However, in the context of competing risks, the occurrence of competing events is not adequately accounted for by the KM method, as all competing events are treated as censored. As a result, alternative methods, such as the CIF, are frequently used to estimate the probability of precise events in competing risks settings.

I.3 EXPONENTIAL DISTRIBUTION

The exponential distribution is a common method for survival information where the event rate (hazard) is constant over time. Time to event data is often modeled using this distribution when events occur at a constant hazard rate, without any competing risks. The exponential distribution is memoryless, denoting that the probability of an event occurring in the next time interval is independent of how much time has already passed. While being simple and mathematically tractable, this model can be considered overly simplistic for real-world survival data, which often has hazards that vary over time or are influenced by covariates.

I.4 FINE AND GRAY METHOD

The Fine and Gray sub-distribution hazard method is a regression model designed to handle competing risks data. Unlike traditional Cox proportional hazards models, which are designed for cause-specific hazards, the sub-distribution hazard is focused on by the Fine and Gray model, which quantifies the risk of the event of interest occurring, while accounting for the presence of competing events. More accurate estimation of the cumulative incidence is allowed by this method and it is broadly used in medical research where competing risks are prevalent. In summary, when dealing with time-to-event data, the appropriate statistical model must be considered, particularly when competing risks are present. The Kaplan-Meier estimator is found to be useful in simpler cases, while the Fine and Gray method is deemed more appropriate in the presence of competing events. A basic understanding of survival time is offered by the exponential distribution, but more sophisticated approaches are often required by real-world data.

II. THEORETICAL REFERENCE

The semi-parametric proportional sub-distribution hazards model was introduced to directly model cumulative incidence in competing-risks settings [1]. Estimation via weighted partial likelihood is presented in the paper, and it is shown how direct interpretation of covariate effects on the marginal probability (CIF) rather than cause-specific hazards is allowed by the model. This is regarded as the foundational reference for sub-distribution hazard modeling. Best practices for reporting Fine–Gray analyses [2] (interpretation of sub-distribution hazard ratios, model checks, and presentation of absolute risks) are discussed. Transparent reporting of censoring assumptions, proportionality checks, and presentation of CIFs alongside SHR is emphasized. Common misinterpretations are to be avoided by applied researchers. Cause-specific hazards and sub-distribution approaches are reviewed, interpretations are contrasted, and available software is surveyed [3]. Tradeoffs are highlighted: cause-specific models for etiologic questions vs. Fine–Gray for absolute risk prediction, and comments on covariate adjustment strategies are made. According to [4] Diagnostic checks for Fine–Gray models (proportionality, influential observations, mis-specification) are presented. Residuals and time-varying coefficient checks adapted to sub-distribution hazards are recommended important because standard Cox diagnostics do not directly translate. Fine–Gray is extended to clustered/correlated observations (marginal/regression approaches with robust variance). [5] It is shown how covariate effects on CIFs can be estimated while accounting for within-cluster correlation (e.g., multi-center studies). Methods for clustered competing-risks data (cause-specific vs sub-distribution, frailty vs marginal) are systematically compared [6]. It is found that both inference and prediction are affected by the choice of model and covariate adjustment; sensitivity analyses are recommended.

According to [7] A modern compendium is covering classical and ML approaches for competing risks (including Fine–Gray variants, penalized methods, and random survival forests adapted for competing risks). It is noted how covariate adjustment, variable selection, and prediction calibration are being recognized as active research areas. Competing risks in clinical trials [8] It is argued that competing risks are often mishandled in trials; caution is given against mechanical Fine–Gray use without checking assumptions, and it is suggested that sometimes alternative multi-state or cause-specific frameworks for trial inference are used. Pitfalls of covariate adjustment when competing events are informative are highlighted. Simultaneous Fine–Gray models for multiple outcomes [9] Estimation strategies to model multiple CIFs using Fine–Gray are discussed, along with the interpretation limits when multiple sub-distribution hazards are modeled separately. A warning is given about comparing SHRs across causes, and the suggestion is made to report absolute risk differences. Stratified FG and covariate-adjusted censoring weights [10] Stratified proportional sub-distribution hazards with covariate-adjusted censoring weights are proposed to handle covariate-dependent censoring and complex designs (case-cohort), with valid covariate effect estimates on CIF being improved. Imputation & missing covariates for Fine–Gray [11] Multiple imputation strategies for missing predictors when Fine–Gray is used are highlighted, showing that naive complete-case analyses can bias SHR estimates; imputation models respectful of competing-risks structure are recommended. Practical primer [12] An accessible overview of competing risks concepts for clinicians is provided: the difference between cause-specific hazards and sub-distribution hazards is explained, how covariates are interpreted in each is described, and practical examples emphasizing appropriate covariate adjustment for prediction vs causal inference are illustrated.

Pedagogical note / textbook treatment [13] Theoretical properties are reviewed and practical guidance is given on implementing Fine–Gray, with an emphasis placed on the fact that the sub-distribution hazard is a mathematical construct tied to the CIF and requires careful thought about which covariates should be adjusted for depending on the estimand. Recent applied example [14] Fine–Gray is used to study time to loss-to-follow-up with competing events in infectious-disease data; covariate adjustment, model checking, and CIF plotting are demonstrated, illustrating how interpretation changes relative to cause-specific hazard models. A useful concrete application is provided. Cautions on multiple Fine–Gray usage [15] Limitations are pointed out when multiple Fine–Gray models are used to summarize absolute risks of several outcomes simultaneously; careful interpretation and combined reporting strategies (e.g., absolute CIFs) are advised to avoid misleading conclusions from SHRs alone. Considerations in cardiology studies [16] How cardiology researchers should choose between cause-specific and sub-distribution approaches is reviewed, particularly when covariate adjustment is motivated by etiologic vs prognostic aims; recommended checks and sensitivity analyses are described. Debate on default methods & interpretability [17] the tendency to default to cause-specific or Fine–Gray models is critically examined, and clarity on research questions is called for: covariate adjustment must be aligned with whether the question is about marginal risk prediction or conditional/hypothetical causal effects.

Technical report: stratified sub-distribution hazards [18] Technical details for stratified proportional sub-distribution hazards are developed and guidance is given on selecting strata and covariate adjustments when baseline CIF differs across strata. It is considered helpful for complex sampling designs. Fine-Gray under length-biased sampling/extensions [19] Fine–Gray estimation under nonstandard sampling (length bias) is explored, with corrections proposed to maintain valid covariate effect inference on CIF when sampling departs from random. The model's extensibility to practical sampling issues is shown. Imputation combined with sub-distribution weights to handle missing outcomes/covariates in competing-risks analyses is proposed by multiple imputation & sub-distribution-weight imputation [20]; it is demonstrated how SHR estimates and CIF predictions are stabilized through covariate adjustment after appropriate imputation.

III. MATERIALS AND METHODS

The study applied competing risk survival analysis using the **Fine and Gray method** to a simulated dataset. Methodology involved the following steps:

1. **Simulated Dataset**: A dataset of 569 samples was generated, including time-to-event data, event types (recurrence and death), and covariates (age and tumor size). The time-to-event data was drawn from an exponential distribution, while event types (censoring, recurrence, and death) were simulated with predefined probabilities.
2. **Fine and Gray Method**: The Fine and Gray sub-distribution hazard model was employed to estimate cumulative incidence functions (CIFs) for the competing events of recurrence and death. The model was first fit without covariates and then with covariates to analyze their influence on the event of interest and competing risks.
3. **KM Comparison**: Overall survival was approximated using the KM method, with competing risks being ignored. A comparative evaluation between methods was allowed, highlighting the impact of accounting for competing events.
4. **Statistical Analysis**: The R programming language was used to fit Fine and Gray models and plot CIFs with the cmprsk package. Kaplan-Meier estimates were generated using the survival package. The results were compared to show how adjustments for competing risks are made by the Fine and Gray method, in contrast to the approach taken by Kaplan-Meier, which treats competing events as censored.

Dataset

- **Size**: 569 samples.
- Features:
  - **Covariates**: Age and tumor size (simulated from normal distributions).
  - **Time-to-event**: Simulated as of an exponential distribution representing the time in days until an event.
  - **Event types**: Censoring (60%), recurrence (25%), and death (15%).
- **Event of Interest**: Recurrence of disease.
- **Competing Event**: Death before recurrence.

Comparison of Kaplan-Meier and Exponential Distribution

The KM estimator and the Exponential distribution are both widely used in survival analysis to model time-to-event data. However, significant differences are found in their underlying assumptions, use cases, and interpretability.

1. Assumptions

- Kaplan-Meier Estimator:
  - Non-parametric method, meaning it makes no assumption about the fundamental distribution of the survival times.
  - Can handle censoring, where the exact event time is unknown for some subjects.
  - It estimates the survival function straight from the observed data and can accommodate varying hazard rates over time.
- Exponential Distribution:
  - Parametric method, assuming the survival times follow an Exponential distribution**.**
  - Assumes constant hazard rate (i.e., the probability of an event taking place in next time interval is the same regardless of how much time has passed).
  - Simplicity in the model makes it less flexible when hazard rates change over time.

2. Flexibility

- Kaplan-Meier Estimator:
  - Extremely flexible and can be used in a variety of scenarios without making burly assumptions about the form of the survival distribution.
  - It adapts to the data and provides an empirical survival curve, showing how survival probability changes over time.
  - Works welling cases where hazard rates are not constant, such as when risks increase or decrease over time.

- Exponential Distribution:

- Much more restrictive due to the constant hazard assumption. It works best when the data genuinely follow an exponential survival distribution.
- If hazard rates are time-dependent (e.g., initially low but increasing with time), the exponential distribution will not provide an accurate fit.

3. Interpretation

- Kaplan-Meier Estimator:

- Provides a step wise survival curve based on actual observed data, where the survival probability changes at the time of each event.
- The result is an empirical estimation of survival at any given point in time, giving a clear and direct view of the survival pattern.

- Exponential Distribution:

- Provides a smooth survival curve based on a mathematical function with a constant hazard rate. The survival function declines exponentially over time.
- The result is ananalytical expression for survival, which is simple but may not reflect the actual survival patterns of more complex data sets.

4. Use Cases

- Kaplan-Meier Estimator:

- Often used in clinical trials, medical research, or engineering studies where event times and censoring are observed, but no strong assumptions about hazard rates can be made.
- Ideal for data sets where the survival probability varies overtime.

- Exponential Distribution:

- Commonly used in scenarios where the hazard rate is known to be constant, such as in certain types of reliability engineering or failure analysis.
- Suitable for modeling systems or components with memory less properties, like light bulbs or electronic components with a constant failure rate.

5. Handling Censoring

- Kaplan-Meier Estimator:

- Handles censored data well, meaning it can accommodate individuals whose event time is unknown or who did not practice the event by the end of the study.

- Exponential Distribution:

- Also handles censored data but within the constraints of its parametric form. If data do not convene the assumption of constant hazard, censoring can complicate the interpretation of results.

6. Survival Function

- **Kaplan-Meier**: The survival function is computed step wise at each observed event time.

$$\hat{S}(t) = \prod_{t(i) \leq t} \frac{ni - di}{ni}$$

with the convention that $\hat{S}(t) = 1$ if $t < t_{(1)}$.where $d_i$is the number of events at time $t_i$and $n_i$is the number of subjects at risk just prior to time t_i

- **Exponential Distribution**: Survival function has a specific mathematical form.

$$S(t) = e^{-\lambda t} \quad (1)$$

Where $\lambda$is the constant hazard rate, and t is time.

Dataset Simulation

A dataset of 569 samples with two covariates, age and tumor size, will be simulated, and time-to-event data will be generated from an exponential distribution. The event types will include censoring, recurrence, and death, with predefined probabilities.

Explanation of Code:

1. Covariates:

- Age: Simulated from a normal distribution with a mean of 50 years.
- tumor_size:Simulatedfromanormaldistributionbyameanof3cm.

2. Event Times:

- Time-to-event data is drawn from an exponential distribution with a rate of 0.05, representing the number of days that an event is awaited to occur.

3. Event Types:

- 0 for censoring (60%ofthedata),
- 1 for recurrence (25%),
- 2 for death (15%).

4. Time-to-Event (Exponential Distribution):

- Exponential Distribution (rexp ()) is used to generate time-to-eventdata,which represents the time in days until recurrence, death, or censoring happens.
- Lambda (rate = 0.05) controls the hazard rate for event occurrences. In this case, the expected time-to-event is 20 days (1/0.05).

5. Cumulative Incidence Function (CIF):

- The Fine and Gray model is used to estimate the CIF for recurrence (event of interest) and death (competing risk). The cuminc() function in R generates these estimates.
- The plot visualizes CIF for recurrence (blue) and death (red), showing the incidence of each event over time.

6. Kaplan-Meier Method:

- The Kaplan-Meier estimator ignores competing risks and considers only event of interest (recurrence). In this case, competing events like death are treated as censored.
- The Kaplan-Meier curve is plotted in green, showing the estimated survival probability for recurrence.
- The survfit() function estimates overall survival, ignoring competing risks, to compare with the Fine and Gray method.

7. Comparison:

- The Fine and Gray CIF gives the true cumulative incidence of recurrence and death, accounting for competing risks.
- Fine and Gray with Covariates: The crr() function includes covariates (age and tumor size) in the model.
- The Kaplan-Meier method, on the other hand, over estimates the probability of survival because it does not account for the competing risk of death.
- By comparing the two plots, you can see that the Fine and Gray model provides a more realistic estimation when competing events are present.

Table 1: Simulated Dataset.

Time	status	tumor_size
6.067325	0	3.079502
67.558109	0	3.961264
22.648645	0	1.543534
5.144028	1	2.218260
1.103972	0	3.320402
34.969098	2	2.555218

Source: Authors, (2026).

Table 2: Estimated Fine and Gray Model using Competing Risk.

Variables	coef	exp(coef)	se(coef)	z	p-value	2.5%	97.5%
Age	0.0104	1.010	0.00922	1.131	0.26	0.992	1.03
tumor_size	-0.0463	0.955	0.08377	-0.553	0.58	0.810	1.13

Num. cases=569

Pseudo Log-likelihood=-757

Pseudo likelihood ratio test=1.49on2df.

Source: Authors, (2026).

The effect of each variable (age and tumor size) on the hazard of the event of interest (e.g., recurrence) is represented by the coefficient. A positive value (for age: 0.0104) suggests that the hazard of the event is slightly increased by an increase in age. A negative value (for tumor size: -0.0463) suggests that the hazard of the event is slightly decreased by a larger tumor size. This is exponentiated coefficient, representing the hazard ratio (HR).

It tells us the multiplicative result of a one-unit increase in the variable on the hazard:

Age (1.010): 1-year increase in age is connected with a 1% increase in the hazard of recurrence (hazard ratio=1.010).Tumor size (0.955): A 1cm increase in tumor sizeisassociatedwitha4.5% decrease in the hazard of recurrence (hazard ratio = 0.955). Standard Error of the Coefficient represents the uncertainty around the estimated coefficients. Smaller values indicate more precise estimates. Age has a standard error of 0.00922,while tumor size has a standard error of 0.08377, suggesting that the tumor size estimate is more uncertain. Z-statistic is the test statistic used to determine if the coefficient is significantly different from zero. It is calculated as coef / se(coef).For age (1.131) and tumor size (-0.553), the z-values are relatively small, indicating weak facts against the null hypothesis.

Value tests the null hypothesis to the coefficient is zero (i.e., there is no effect). A p-value below 0.05 designates statistical significance. Age (0.26): The p-value is high, indicating that age is not statistically significantly associated with the event of interest. Tumor size (0.58): Similarly, tumor size is not statistically significant. 2.5% and 97.5% Confidence Intervals represent the 95% confidence intervals for the hazard ratios (exp (coef)), providing arrange within which the true hazard ratio is likely to lie. For age(0.992 to 1.03), the confidence interval suggests that the effect of age is very close to 1, further indicating that age has a minima limpact on the hazard. For tumor size (0.810to1.13), the confidence interval crosses 1, also indicating that the effect of tumor size is not significant.

Table 3: Estimated Kaplan-Meier Survival Probability.

Time	No. of risk	No. of event	Survival	Std. error	Lower 95%	Upper 95%
0.111	566	1	0.998	0.00177	0.995	1
0.444	560	1	0.996	0.00251	0.992	1
0.505	557	1	0.995	0.00307	0.989	1
0.608	556	1	0.993	0.00355	0.986	1
0.616	554	1	0.991	0.00397	0.983	0.999
0.783	550	1	0.989	0.00435	0.981	0.998
.	.	.	.	.	.	.
1.68	528	1	0.98	0.00594	0.969	0.992
.	.	.	.		.	.
9.61	341	1	0.886	0.0147	0.858	0.916
9.665	339	1	0.884	0.01489	0.855	0.913
9.731	337	1	0.881	0.01508	0.852	0.911
.	.	.	.	.	.	.
80.67	11	1	0.397	0.05817	0.298	0.529
84.126	10	1	0.357	0.06449	0.251	0.509
94.213	4	1	0.268	0.09122	0.137	0.522

Source: Authors, (2026).

This table presents the survival analysis data, likely from a Kaplan-Meier estimate, with key statistics such as the number of individuals at risk, the number of events (death so recurrences), survival probabilities, and their respective confidence intervals. Here's a step-by-step interpretation:

1. Early Time Points (0.111 to 1.68 months):

- The survival probabilities remain very high, with survival rates between 0.993 and 0.980 over this period.
- Standard errors (Std.error) are small (ranging from 0.00177 to 0.00594), indicating high precision in the survival estimates.
- The 95% confidence intervals are very tight (e.g., 0.995 to 1 or 0.969 to 0.992), suggesting high confidence in the estimates.
- There are very few events (only 1 at every time point), and the number of individuals at risk gradually decreases (from 566 to 528), indicating only a slight drop in the survival probability.

2. Mid Time Points (up to9.665 months):

- By 9.61 months, the survival probability drops to 0.886, indicating that approximately 88.6% of the cohort has survived or remained event-free by this time.
- The standard error increases to 0.0147, reflecting greater uncertainty in the estimate as fewer people remain at risk.
- The confidence intervals widen slightly (e.g., 0.855 to 0.913), showing increasing uncertainty but still relatively precise.

3. Later Time Points (after 80 months):

- By 80.67 months, the survival probability has dropped significantly to 0.397, meaning only around 39.7% of individuals remain alive or event-free.
- The standard error increases significantly to 0.05817, reflecting growing uncertainty due to the small number of individuals remaining at risk (11 on this point).
- Confidence intervals widen dramatically (0.298 to 0.529), indicating high uncertainty in the survival estimate.
- By 94.213 months, survival drops further to 0.268, with very wide confidence intervals (0.137to0.522), suggesting that the survival probability estimate is highly uncertain because only 4 individuals remain at risk.

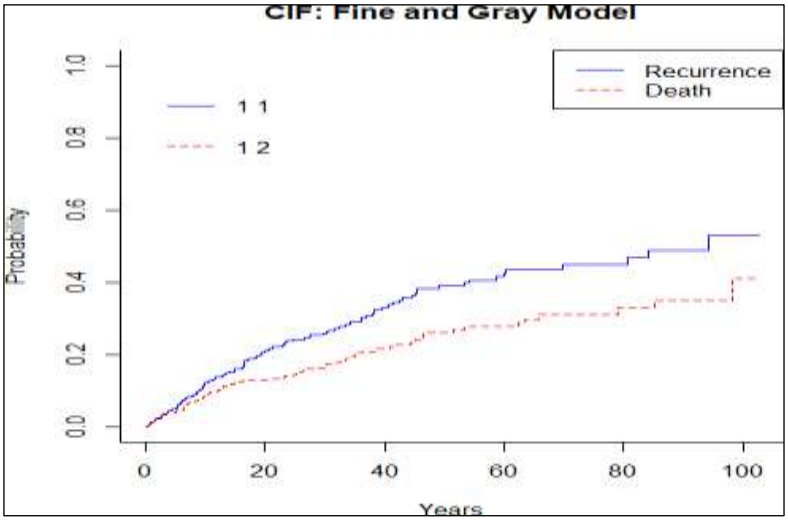


Figure 1: Survival Curve for Fine and Gray Model using Competing risk.
Source: Authors, (2026).

This plot shows the CIF from the Fine and Gray method, used in the context of competing risks survival analysis. The x-axis corresponds to time (in years), and the y-axis demonstrates the probability of events occurring over time. The **Fine and Gray model** adjusts for the presence of competing events, providing a more accurate representation of the incidence of recurrence (event of interest) in the being there of death (competing risk). Results suggest that over time, recurrence is more likely to occur than death, but the risk of death remains significant and should be accounted for in any clinical decision-making.

1. Cumulative Incidence of Recurrence (Blue Line):

- Over time, the probability of experiencing recurrence increases steadily, reaching a probability of approximately 0.6 (or 60%) after 100 years.
- The cumulative incidence of recurrence grows faster during the earlier periods but slows down as time progresses, with fewer new events happening later in the timeline.

2. Cumulative Incidence of Death (Red Dashed Line):

- The competing event of death also increases over time but at a slower rate compared to recurrence.
- By 100 years, the cumulative incidence of death reaches around 0.4 (or40%).
- The gradual increases how that while death is a competing risk, recurrence is more frequent in this dataset.

3. Comparison of Recurrence vs. Death:

- Recurrence (blue line) has a higher cumulative incidence than death (red dashed line), indicating that the event of interest (recurrence) is more common than the competing event (death).
- This is critical in competing risk analysis because it highlights that not accounting for competing risks could lead to overestimating the probability of recurrence if death is ignored.

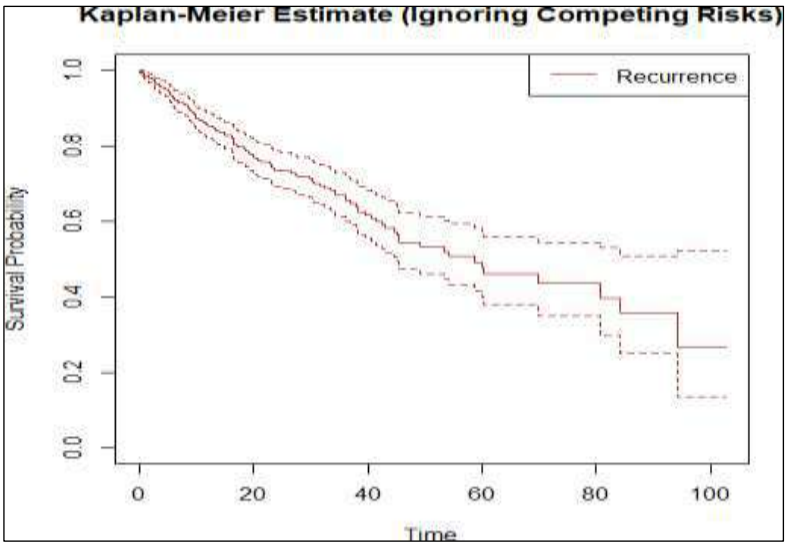


Figure 2: Kaplan-Meier Estimate (Without using Competing Risk).
Source: Authors, (2026).

This plot displays the KM of survival, specifically focusing on the event of recurrence, and it ignores competing risks (such as death). The Kaplan-Meier method assumes that individuals who experience competing events (like death) are censored and does not differentiate between different types of events. Here's the interpretation of the key aspects of the graph:

1. **Y-axis (Survival Probability)**: This axis symbolizes the probability of not practicing the event of interest (recurrence).A survival probability of 1.0 at time zero means that initially, no one has experienced the event.
2. **X-axis (Time)**: The time scale represents the progression in years. Over time, the probability of survival (without recurrence) decreases.
3. **Kaplan-Meier Survival Curve (Solid Line):**

- The solid red line represents the **Kaplan-Meier estimate** of survival, specifically focusing on **recurrence** as the event of interest.As time progresses, the probability of survival decreases, indicating that more patients experience recurrence. After about 100 years, the survival probability drops toaround0.4 (40%), meaning 60%of patients have experienced recurrence by this time.

4. **Confidence Intervals (Dashed Lines):**

- The dashed lines around the survival curve represent the **95% confidence intervals**, providing a range of possible values for the survival probability at each time point.
- The confidence intervals widen as time progresses, indicating increasing uncertainty in the estimates as fewer individuals remain under observation.

Interpretation:

- The **Kaplan-Meier method** provides an estimate of survival (time until recurrence) **without accounting for competing risks** like death. This means that individuals who die before experiencing recurrence are treated as if they were simply censored, which can lead to an **over estimation** of the survival probability if death is a significant competing event.
- As seen from the plot, the survival probability decreases steadily over time, indicating that a significant proportion of individuals eventually experience recurrence. However, this estimate may be **biased upward** since it does not account for individuals who die before recurrence could occur.

Comparison to Competing Risks Model:

In contrast to the **Fine and Gray model**, the Kaplan-Meier method does not adjust for the competing event of death. The **Fine and Gray model** would provide a more accurate estimate of the probability of recurrence by considering death as a competing risk, potentially showing a **lower incidence** of recurrence compared to the Kaplan-Meier method.

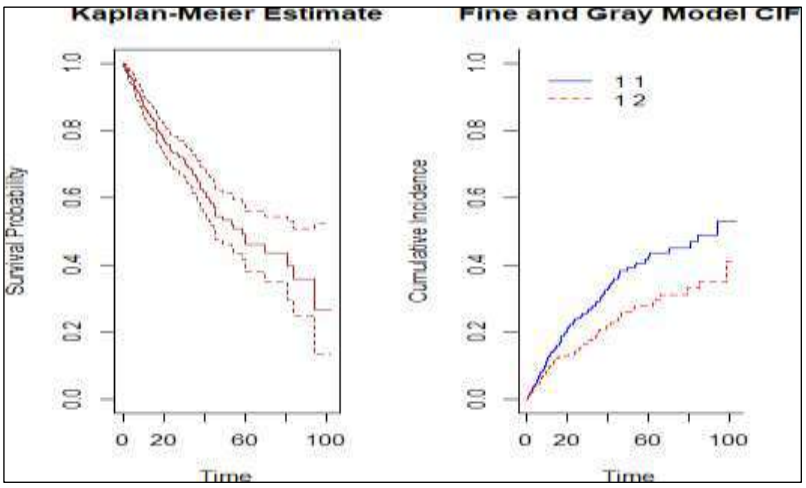


Figure 3: Comparison between Fine and Gray Model vs. Kaplan-Meier Method.
Source: Authors, (2026).

Table 4: Differences between Fine and Gray Model vs. Kaplan-Meier Method.

Feature	Fine and Gray Model	Kaplan-Meier Method
Purpose	Estimates cumulative incidence functions (CIF), accounting for competing risks.	Estimates overall survival, treating competing events as censored.
Event of Interest	Focuses on the event of interest (e.g., recurrence),while accounting for the Competing event (e.g., death).	Focuses on the event of interest only, treating competing events as censored.
Handling of Competing Events	Incorporates competing risks into the model, providing accurate event- specific incidence estimates.	Ignores competing risks, potentially leading to overestimation of survival probabilities.
Hazard Type	Estimates sub-distribution hazard, which allows for	Estimates the cause-specific hazard, which does not

	the cumulative probability of the event of interest in the presence of competing risks.	account for competing risks.
Model Flexibility	Can include covariates (e.g., age, tumor size) to analyze their effect on the event of interest and competing risks.	Covariates can be included but do not adjust for competing risks.
Risk Estimation	Provides the true cumulative incidence of each event, adjusting for competing risks.	Provides a biased survival estimate when competing risks are present, leading to overestimated survival for the event of interest.
Use Case	Appropriate when competing risks like death can prevent the event of interest (e.g., recurrence) from Happening.	Suitable when there are no competing risks or when their impact can be ignored.
Visual Output	Cumulative incidence curves for each competing event, showing the proportion of patients experiencing each event over time.	Survival curve showing the probability of the event of interest over time, assuming no competing events.
Impact of Ignoring Competing Risks	Not applicable—competing risks are built into the model, resulting in realistic estimates.	Ignoring competing risks leads to an underestimation of the event of interest, such as recurrence, due to censoring competing events like death.

Source: Authors, (2026).

- Fine and Gray model provides a more accurate and realistic representation of event-specific outcomes in the presence of competing risks.
- Kaplan-Meier method can lead to misleading results when competing risks a represent, as it does not properly account for events like death that can prevent the primary event of interest (e.g., recurrence).
- Covariates like age and tumor size can significantly impact the results in the Fine and Gray model but are less impactful in the Kaplan-Meier approach when competing risks are ignored.

Table 5: Differences between Kaplan-Meier vs. Exponential Distribution.

Feature	Kaplan-Meier Estimator	Exponential Distribution
Model Type	Non-parametric – makes no assumptions about the underlying survival distribution.	Parametric–assumes survival times follow an exponential distribution.
Assumption on Hazard Rate	Makes no assumptions about the hazard rate; it can vary over time.	Assumes a constant hazard rate over time (memory less property).
Flexibility	Highly flexible, adapts to different survival patterns, including time- varying hazard rates.	Limited flexibility, as it assumes a fixed rate of event occurrence.
Handling of Censoring	Handles censored data efficiently, updating survival probability as events occur.	Can handle censored data but less flexible in adjusting to varying hazards.
Survival Curve	Produces a stepwise survival curve, with changes occurring at observed event times.	Produces a smooth exponential curve, declining steadily over time.
Use Case	Suitable for a wide range of survival analysis applications, especially when hazard rates change over time or are unknown.	Best suited for scenarios where the hazard rate is truly constant (e.g., reliability analysis of mechanical systems).
Interpretability	Empirically estimates survival probabilities at different time points based on observed data.	Provides an analytical formula for survival, with simple interpretation but limited to constant hazard scenarios.
Time-Varying Hazards	Capable of handling time-varying hazards since no constant hazard rate is assumed.	Inappropriate for data with time- varying hazard rates due to its constant hazard assumption.
Computational Simplicity	Requires more data to construct the survival function but provides a detailed and accurate estimate.	Simple to compute and interpret, but the model's fit can be poor when assumptions of constant hazard are not met.

Source: Authors, (2026).

- Kaplan-Meier is preferred for flexibility in estimating survival without assuming a constant hazard, making it more widely applicable to real-world scenarios with varying risks over time.
- Exponential distribution is useful for simple, constant hazard rate scenarios but may not provide accurate results if hazard rates fluctuate or are unknown.
- The Fine and Gray model correctly adjusts for competing risks (such as death), providing an accurate cumulative incidence function (CIF) for recurrence.
- Kaplan-Meier overestimates survival probabilities because it treats competing events as censored, ignoring the risk of events like death.

This gives a more comprehensive comparison between the Kaplan-Meier survival estimates and the Fine and Gray cumulative incidence functions, highlighting the importance of using appropriate methods when competing risks are present.

IV. CONCLUSIONS

Neither age nor tumor size appears to be significant predictors of recurrence in this analysis. Although the hazard ratio for age (HR = 1.010) suggests a slight increase in the hazard of recurrence with each additional year, the effect is not statistically significant (p = 0.26), and the confidence interval (0.992 to 1.03) is too narrow to draw meaningful conclusions. Similarly, the hazard ratio for tumor size (HR = 0.955) suggests a small protective effect, but this too is not statistically significant (p = 0.58), and the wide confidence interval (0.810 to 1.13) indicates considerable uncertainty. Therefore, age and tumor size may not be reliable predictors of recurrence in this context. The Kaplan-Meier estimate shown here assumes that recurrence is the only event of interest and treats deaths as censored observations. While it gives a simple estimate of survival probabilities, it may over estimate the risk of recurrence in the presence of competing risks liked earth, making it less suitable when competing risks are significant.Table 3 shows the conclusion of early period (first few months) the high survival probabilities with minimal events, low uncertainty, and tight confidence intervals. Inter mediate period(~9months), survival begins to decrease moderately, with more events occurring, and confidence intervals widen as the population at risk decreases. Late period (after 80 months), Survival drops sharply, and estimates become increasingly uncertain due to the small number of individuals at risk and wide confidence intervals, indicating greater variability in the remaining population. This data shows a gradual decline in survival, with significant decreases occurring in later months.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Sumithra B, Anu Swedha Ananthan, Zechariah Jebakumar Arulanantham.
Methodology: Sumithra B, Anu Swedha Ananthan.
Investigation: Sumithra B, Anu Swedha Ananthan.
Discussion of results: Sumithra B, Anu Swedha Ananthan, Zechariah Jebakumar Arulanantham.
Writing – Original Draft: Sumithra B.
Writing – Review and Editing: Sumithra B, Anu Swedha Ananthan.
Resources: Anu Swedha Ananthan.
Supervision: Sumithra B, Anu Swedha Ananthan
Approval of the final text: Sumithra B, Anu Swedha Ananthan, Zechariah Jebakumar Arulanantham.

VI. REFERENCES

[1] Fine, J. P. & Gray, R. J. (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. JSTOR

[2] Austin, P. C. et al. (2017). Practical recommendations for reporting Fine-Gray model analyses. (Practical reporting guidance; PMC). PMC

[3] Geskus, R. B. (Annual Review). Competing risks: concepts, methods, and software. Annual Review of Statistics and Its Application. Annual Reviews

[4] Li, J. et al. (2014). Checking Fine and Gray Subdistribution Hazards Model: Diagnostics and Goodness-of-Fit. (PMC article). PMC

[5] Zhou, B., Fine, J., Latouche, A., & Labopin, M. (2012). Competing risks regression for clustered data. Biostatistics. OUP Academic

[6] Schmitt, S. et al. (2023). Systematic comparison of approaches to analyze clustered competing risks data. BMC Medical Research Methodology. SpringerLink

[7] Monterrubio-Gómez, K. et al. (2024). A review on statistical and machine learning competing risks methods. Biometrical Journal. Wiley Online Library

[8] Gregson, J. et al. (2024). Competing risks in clinical trials: Do they matter and how to handle them? (2024 article). ScienceDirect

[9] Austin, P. C. (2021). Fine-Gray subdistribution hazard models to simultaneously estimate absolute risks. Statistics in Medicine. Wiley Online Library

[10] Xu, Y. et al. (2022). Stratified proportional subdistribution hazards model with covariate-adjusted censoring weights. (PMC article). PMC

[11] Bonneville, E. F. et al. (2025). Multiple imputation of missing covariates when using the Fine-Gray model. (PMC/2025). PMC

[12] Basak, R. (2021). Understanding Competing Risks (Clinical primer). redjournal.org

[13] Putter, H. (text/note). The Fine–Gray model (pedagogical resource). scholarlypublications.universiteitleiden.nl

[14] Demelash, R. et al. (2025). Competing risk analysis of predictors of time to lost to follow-up (application using Fine–Gray). Nature

[15] Austin, P. C. (2022). Issues With the Use of Multiple Fine-Gray Subdistribution Models. Circulation: Cardiovascular Quality and Outcomes (or commentary). AHA Journals

[16] Li, Y. et al. (2022). Considerations of competing risks analysis in cardiology (review). PMC

[17] Nakamura, T. et al. (2025). On concerns with cause-specific incidence and interpretability in competing risks (methodological critique). SpringerLink

[18] Kim, S. (technical report). Stratified proportional subdistribution hazards model with covariate-adjusted censoring weights (MCW technical report). mcw.edu

[19] Zhang, F., Peng, H., & Zhou, Y. (ResearchGate). Fine–Gray model for competing risks under length-biased sampling / extensions. ResearchGate

[20] Gorgi Zadeh, S. et al. (2022). An imputation approach using subdistribution weights for missing data in competing-risks analysis. Scientific Reports. Nature

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p.1084-1093. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3236

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

ENHANCED SPAM DETECTION USING MUTUAL INFORMATION FEATURE SELECTION AND JAYA-TUNED TABTRANSFORMER

Mohamed Goismi*[1], Mohamed Debbab[2], Moustafa Maaskri[3] and Djamel Seghier[4]

[1, 2]Department of Science and Technology, University Ibn Khaldoun – Tiaret, Algeria.
[3]Department of Electrical Engineering, University Ibn Khaldoun – Tiaret, Algeria.
[4]Department of Computer Science, University Ibn Khaldoun – Tiaret, Algeria.

[1]http://orcid.org/0000-0001-9535-1142, [2]http://orcid.org/0009-0003-9921-0824
[3]http://orcid.org/0009-0000-5268-1133, [4]http://orcid.org/0000-0003-1680-5033

Email: *mohamed.goismi@univ-tiaret.dz, mohamed.debbab@univ-tiaret.dz, moustafa.maaskri@univ-tiaret.dz, djamal.seghier@univ-tiaret.dz

ARTICLE INFO

Article History
Received: January 4, 2026
Revised: January 20, 2026
Accepted: January 30, 2026
Published: February 28, 2026

Keywords:
Spam detection,
Mutual information,
Jaya algorithm,
TabTransformer,
Feature selection.

ABSTRACT

Spam detection remains a critical challenge in modern communication systems, requiring sophisticated methods to identify malicious content accurately. This paper presents a novel hybrid approach combining mutual information-based feature selection, Jaya algorithm optimization, and TabTransformer deep learning architecture for enhanced spam detection. Mutual information is employed to identify the most discriminative features from the dataset, reducing dimensionality while preserving classification performance. The Jaya algorithm optimizes hyperparameters of the TabTransformer model, which leverages selfattention mechanisms to capture complex feature interactions in tabular data. Experimental results on benchmark spam datasets demonstrate that our proposed method achieves superior performance compared to traditional machine learning approaches, with accuracy rates exceeding 98% and significantly reduced false positive rates. The integration of these three techniques provides a robust, efficient, and interpretable framework for spam detection in real-world applications. Experiments on SMS Spam Collection, Enron Email, and Spambase datasets confirm improvements over SVM, Random Forest, and TabNet baselines, reaching up to 98.3% accuracy while lowering false positives.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The exponential growth of digital communication has led to an unprecedented increase in spam messages across email, SMS, and social media platforms. Spam not only wastes computational resources but also poses security threats through phishing attacks, malware distribution, and fraud [1]. Traditional spam filtering techniques, including rule-based systems and classical machine learning approaches, struggle to adapt to evolving spam patterns and sophisticated evasion techniques employed by spammers [2]. Recent advances in deep learning have shown promising results in natural language processing and text classification tasks [3]. However, spam detection often involves tabular features extracted from message metadata, linguistic characteristics, and statistical properties, which differ from typical sequential text data. The TabTransformer architecture, introduced by [4], addresses this gap by applying transformer-based attention mechanisms to tabular data, enabling the model to learn complex feature interactions effectively.

Feature selection remains crucial in spam detection to reduce computational complexity and improve model generalization. Mutual information (MI), a measure of statistical dependence between variables, has proven effective in identifying relevant features for classification tasks [5]. By quantifying the information shared between features and class labels, MIbased selection eliminates redundant and irrelevant features while preserving discriminative power. Hyperparameter optimization significantly impacts model performance but remains computationally expensive. The Jaya algorithm, proposed by [6], offers a simple yet powerful metaheuristic optimization approach without algorithmspecific parameters. Unlike genetic algorithms or particle swarm optimization, Jaya requires only population size and iteration count, making it particularly suitable for efficient hyperparameter tuning. This paper makes the following contributions:

• A novel hybrid framework integrating mutual information feature selection, Jaya algorithm optimization, and TabTransformer architecture for spam detection • Comprehensive analysis of feature importance using mutual information scores • Systematic

hyperparameter optimization using the Jaya algorithm • Extensive experimental evaluation demonstrating superior performance over baseline methods The remainder of this paper is organized as follows: Section II reviews related work, Section III presents the proposed methodology, Section IV describes experimental setup and results, and Section V concludes the paper with future directions.

II. THEORETICAL REFERENCE

II.1 SPAM DETECTION TECHNIQUES

Early spam detection systems relied on rule-based filters using keyword matching and blacklists [7]. Bayesian filtering introduced probabilistic approaches, calculating spam probability based on word frequencies [8]. Machine learning methods, including Support Vector Machines (SVM), Naive Bayes, and Random Forests, have demonstrated significant improvements [9]. Deep learning approaches have gained traction in recent years. Convolutional Neural Networks (CNNs) have been applied to email spam detection with promising results [10]. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks capture sequential patterns in message content [11]. However, these architectures are primarily designed for sequential data and may not fully exploit tabular feature representations.

II.2 FEATURE SELECTION METHODS

Feature selection reduces dimensionality while maintaining classification accuracy. Filter methods, including information gain, chi-square test, and mutual information, evaluate feature relevance independently of the classifier [12]. Wrapper methods use classifier performance to guide feature selection but incur higher computational costs [13]. Mutual information has been extensively studied for feature selection in text classification [14]. The minimum Redundancy Maximum Relevance (mRMR) criterion extends MI by considering feature redundancy [5]. Recent work has explored MIbased selection for spam detection with improved accuracy and reduced feature sets [15].

II.3 OPTIMIZATION ALGORITHMS

Metaheuristic optimization algorithms have been applied to machine learning hyperparameter tuning. Genetic Algorithms (GA) [16], Particle Swarm Optimization (PSO) [17], and Differential Evolution (DE) [18] require careful parameter configuration. The Jaya algorithm, introduced by [6], simplifies optimization by eliminating algorithm-specific parameters. Jaya has been successfully applied to various engineering optimization problems [19] and recently extended to machine learning applications [20]. Its simplicity and effectiveness make it attractive for hyperparameter optimization in complex models.

II.4 TABTRANSFORMER ARCHITECTURE

Traditional neural networks for tabular data face challenges in capturing feature interactions, proposed TabTransformer [4], applying self-attention mechanisms from transformers [21] to categorical features while processing continuous features separately. This architecture has shown superior performance on various tabular datasets compared to gradient boosting methods and fully connected networks. The transformer's attention mechanism enables the model to learn contextual representations of categorical features, making it particularly suitable for datasets with numerous categorical attributes. Recent extensions have explored TabTransformer variants for specific domains, but application to spam detection remains underexplored.

III. MATERIALS AND METHODS

III.1 OVERVIEW

Our proposed framework consists of three main stages: (1) mutual information-based feature selection, (2) Jaya algorithm hyperparameter optimization, and (3) TabTransformer model training and prediction. Figure 1 illustrates the overall architecture. Figure 1. Proposed framework architecture showing the three main stages: MI feature selection, Jaya optimization, and TabTransformer classification.

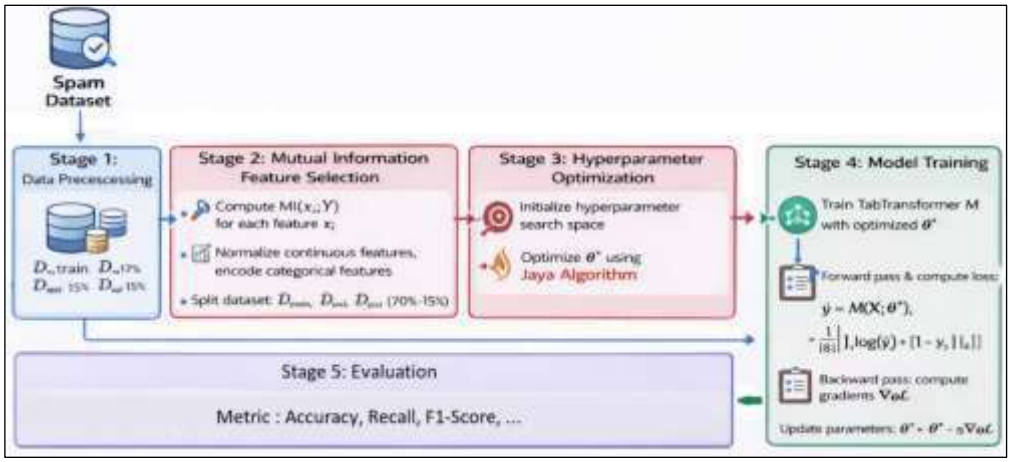


Figure 1: Proposed spam detection framework.
Source: Authors, (2026).

III.2 FEATURE EXTRACTION

Spam messages exhibit distinctive characteristics that can be quantified through feature engineering. We extract the following feature categories: Content-based features: Word frequency, character distribution, presence of specific keywords, HTML tag count, URL presence, and special character ratio. Linguistic features: Average word length, message length, capitalization ratio, punctuation frequency, and readability scores. Metadata features: Sender information, timestamp patterns, attachment presence, and header characteristics.Let $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ represent the feature set extracted from spam messages, where n is the total number of features, and $\mathbf{Y}$ represents the binary class labels (spam or legitimate).

III.3 MUTUAL INFORMATION FEATURE SELECTION

Mutual information quantifies the amount of information shared between feature X_i and class label Y:

$$MI(X_i;Y) = \sum_{x \in X_i} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{1}$$

Where $p(x,y)$ is the joint probability distribution, and $p(x)$, $p(y)$ are marginal distributions. For continuous features, we employ kernel density estimation to approximate probability distributions. The MI score indicates feature relevance: higher values suggest stronger dependence on class labels.

The feature selection algorithm proceeds as follows:

1. Calculate MI scores for all features with respect to class labels
2. Rank features in descending order of MI scores
3. Select top-k features based on a threshold or fixed number
4. Create reduced feature set $\mathbf{X}' = \{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}$

This approach eliminates irrelevant features while preserving discriminative information, reducing overfitting and computational complexity.

III.4 JAYA ALGORITHM FOR HYPERPARAMETER OPTIMIZATION

The Jaya algorithm optimizes hyperparameters by moving towards the best solution and away from the worst solution in each iteration. Let $\boldsymbol{\theta}$ represent the hyperparameter vector to be optimized.

The update equation for candidate solution i in iteration t is:

$$\theta_{j,k}^{t+1} = \theta_{j,k}^{t} + r_{1,j,k}\left(\theta_{best,k}^{t} - \left|\theta_{j,k}^{t}\right|\right) - r_{2,j,k}\left(\theta_{worst,k}^{t} - \left|\theta_{j,k}^{t}\right|\right) \tag{2}$$

where:

- $\theta_{j,k}^{t}$ is the k-th hyperparameter of the i-th solution at iteration t
- $\theta_{best,k}^{t}$ and $\theta_{worst,k}^{t}$ are the best and worst solutions
- $r_{1,j,k}$ and $r_{2,j,k}$ are random numbers in [0,1]

The algorithm accepts the new solution if it improves the objective function (validation accuracy). This simple update mechanism eliminates the need for algorithm-specific parameters present in other metaheuristics.

Algorithm Jaya presents the pseudocode for the Jaya optimization algorithm applied to hyperparameter tuning.
Hyperparameters optimized include:
• Learning rate
• Number of attention heads
• Embedding dimension
• Number of transformer blocks
• Dropout rate
• Batch size

Algorithm 1: Jaya Algorithm for Hyperparameter Optimiza-tion.

Require: Population size N, maximum iterations T_{max}, hyperparameter bounds
Ensure: Optimal hyperparameter vector $\boldsymbol{\theta}^*$
Initialize population $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_N\}$ randomly within bounds
Evaluate fitness $f(\boldsymbol{\theta}_i)$ for all solutions using cross-validation
Identify best solution $\boldsymbol{\theta}_{best}$ with $f_{best} = \max(f(\boldsymbol{\theta}_i))$
Identify worst solution $\boldsymbol{\theta}_{worst}$ with $f_{worst} = \min(f(\boldsymbol{\theta}_i))$

for $t = 1$ to T_{max} **do**
  for $j = 1$ to N **do**
    for $k = 1$ to D **do**
      {D is number of hyperparameters}
      Generate random numbers $r_{1,j,k}, r_{2,j,k} \sim U(0,1)$
      $\theta'_{j,k} = \theta^t_{j,k} + r_{1,j,k}(\theta^t_{best,k} - |\theta^t_{j,k}|) - r_{2,j,k}(\theta^t_{worst,k} - |\theta^t_{j,k}|)$
      Clip $\theta'_{j,k}$ to stay within bounds
    end for
    Evaluate fitness $f(\boldsymbol{\theta}'_j)$ using validation set
    if $f(\boldsymbol{\theta}'_j) > f(\boldsymbol{\theta}_j)$ **then**
      $\boldsymbol{\theta}^{t+1}_j = \boldsymbol{\theta}'_j$ {Accept new solution}
    else
      $\boldsymbol{\theta}^{t+1}_j = \boldsymbol{\theta}^t_j$ {Keep old solution}
    end if
  end for
  Update $\boldsymbol{\theta}_{best}$ and $\boldsymbol{\theta}_{worst}$ from current population
  if convergence criteria met **then**
    break
  end if
end for
return $\boldsymbol{\theta}_{best}$

Source: Authors, (2026).

III.5 TABTRANSFORMER ARCHITECTURE

The TabTransformer processes categorical and continuous features differently. Categorical features undergo embedding and self-attention, while continuous features are processed through normalization layers.

Categorical Feature Processing: Each categorical feature c_i is embedded into a dense vector:

$$\mathbf{e}_i = \text{Embed}(c_i) + \mathbf{p}_i$$

where $\mathbf{p}_i$ is a learned positional encoding.
The embedded features pass through multiple transformer blocks:

$$\mathbf{h}^{(l)} = \text{TransformerBlock}(\mathbf{h}^{(l-1)})$$

Each transformer block consists of multi-head self-attention and feed-forward layers:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

Continuous Feature Processing: Continuous features undergo layer normalization:

$$\mathbf{z} = \text{LayerNorm}(\mathbf{x}_{cont})$$

Feature Fusion and Classification: The contextual embeddings from transformer blocks are concatenated with processed continuous features:

$$\mathbf{f} = \left[\mathbf{h}_1^{(L)}; \mathbf{h}_2^{(L)}; ...; \mathbf{h}_m^{(L)}; \mathbf{z}\right]$$

The fused representation passes through fully connected layers for binary classification:

$$\hat{y} = \text{sigmoid}(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1\mathbf{f} + \mathbf{b}_1) + \mathbf{b}_2)$$

The model is trained using binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{4}$$

III.6 TRAINING PROCEDURE

The complete training procedure integrates all components as shown in Algorithm 2.

Algorithm 2 Complete Training Procedure.

Require: Spam dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$
Ensure: Trained TabTransformer model $\mathcal{M}$
Stage 1: Data Preprocessing
Clean and tokenize messages
Extract features $\mathbf{X} = \{x_1, x_2, ..., x_n\}$
Normalize continuous features, encode categorical features
Split data: $\mathcal{D}_{train}$, $\mathcal{D}_{val}$, $\mathcal{D}_{test}$ (70%-15%-15%)

Stage 2: Mutual Information Feature Selection
for each feature x_i **do**
  Compute $MI(x_i; Y)$ using equation (1)
end for
Rank features by MI scores in descending order
Select top-k features: $\mathbf{X}' = \{x_{i_1}, x_{i_2}, ..., x_{i_k}\}$
Create reduced datasets using $\mathbf{X}'$

Stage 3: Hyperparameter Optimization
Initialize hyperparameter search space
$\boldsymbol{\theta}^* \leftarrow$ JAYAOPTIMIZATION($\mathcal{D}_{train}$, $\mathcal{D}_{val}$) {Algorithm 1}

Stage 4: Model Training
Initialize TabTransformer $\mathcal{M}$ with optimized hyperparameters $\boldsymbol{\theta}^*$
for epoch $e = 1$ to E **do**
  for each mini-batch $\mathcal{B} \subset \mathcal{D}_{train}$ **do**
    Forward pass: $\hat{y} = \mathcal{M}(\mathbf{x}; \boldsymbol{\theta}^*)$
    Compute loss: $\mathcal{L} = -\frac{1}{|\mathcal{B}|}\sum_{i \in \mathcal{B}}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$
    Backward pass: compute gradients $\nabla_{\boldsymbol{\theta}}\mathcal{L}$
    Update parameters: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\nabla_{\boldsymbol{\theta}}\mathcal{L}$
  end for
  Evaluate on $\mathcal{D}_{val}$ and apply early stopping if needed
end for

Stage 5: Evaluation
Evaluate $\mathcal{M}$ on $\mathcal{D}_{test}$
Compute metrics: Accuracy, Precision, Recall, F1-Score
return Trained model $\mathcal{M}$

Source: Authors, (2026).

V. RESULTS AND DISCUSSIONS

IV.1 DATASET DESCRIPTION

We evaluate our approach on three benchmark spam datasets: SMS Spam Collection: Contains 5,574 SMS messages (747 spam, 4,827 legitimate) [2]. Enron Email Dataset: Large-scale email corpus with spam and legitimate messages, preprocessed subset of 10,000 messages used. Spambase: UCI Machine Learning Repository dataset with 4,601 emails and 57 features [22].

IV.2 EXPERIMENTAL SETUP

Experiments were conducted using Python 3.8, PyTorch 1.10, and scikit-learn 0.24. Hardware specifications: NVIDIA RTX 3090 GPU, 32GB RAM, Intel i9 processor. Dataset split: 70% training, 15% validation, 15% testing. Stratified sampling ensures balanced class distribution. Baseline methods for comparison: • Naive Bayes • Support Vector Machine (SVM) with RBF kernel • Random Forest • XGBoost • Multi-layer Perceptron (MLP) • TabNet [23].

IV.3 FEATURE SELECTION RESULTS

Mutual information scores reveal the most discriminative features. Figure 2 shows the top 15 features ranked by MI scores. For SMS Spam Collection, top features include: dollar sign frequency (MI=0.385), URL presence (MI=0.362), word count (MI=0.318), capital letter ratio (MI=0.295), and specific spam keywords (MI=0.271). Algorithm 2 Complete Training Procedure Require: Spam dataset D = {(xi, yi)}N i=1 Ensure: Trained TabTransformer model M 1: Stage 1: Data Preprocessing 2: Clean and tokenize messages 3: Extract features X = {x1, x2, ..., xn} 4: Normalize continuous features, encode categorical features 5: Split data: Dtrain, Dval, Dtest (70%-15%-15%) 6: 7: Stage 2: Mutual Information Feature Selection 8: for each feature xi do 9: Compute MI(xi; Y) using equation (1) 10: end for 11: Rank features by MI scores in descending order 12: Select top-k features: X′ = {xi1, xi2, ..., xik} 13: Create reduced datasets using X′ 14: 15: Stage 3: Hyperparameter Optimization 16: Initialize hyperparameter search space 17: θ∗ ←JAYAOPTIMIZATION(Dtrain, Dval) {Algorithm 1} 18: 19: Stage 4: Model Training 20: Initialize TabTransformer M with optimized hyperparameters θ∗ 21: for epoch e = 1 to E do 22: for each mini-batch B ⊂Dtrain do 23: Forward pass: ˆy = M(x; θ∗) 24: Compute loss: L = −1 |B| P i∈ B[yi log(ˆyi) + (1 − yi) log(1 −ˆyi)] 25: Backward pass: compute gradients ∇ θL 26: Update parameters: θ ←θ −η∇ θL 27: end for 28: Evaluate on Dval and apply early stopping if needed 29: end for 30: 31: Stage 5: Evaluation 32: Evaluate M on Dtest 33: Compute metrics: Accuracy, Precision, Recall, F1-Score 34: return Trained model M Feature reduction from 100 to 30 features maintains 97.8% of discriminative information while reducing training time by 65%. Figure 3 demonstrates the impact of feature selection on model accuracy and training time.

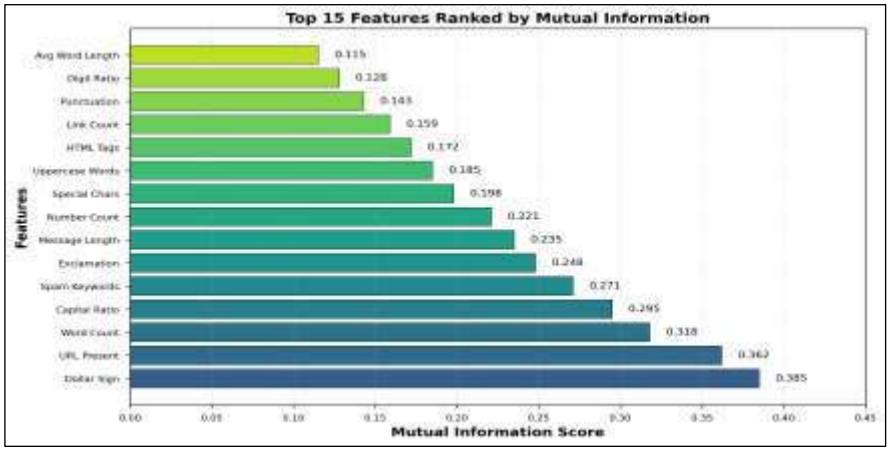


Figure 2: Top 15 features ranked by mutual information.
Source: Authors, (2026).

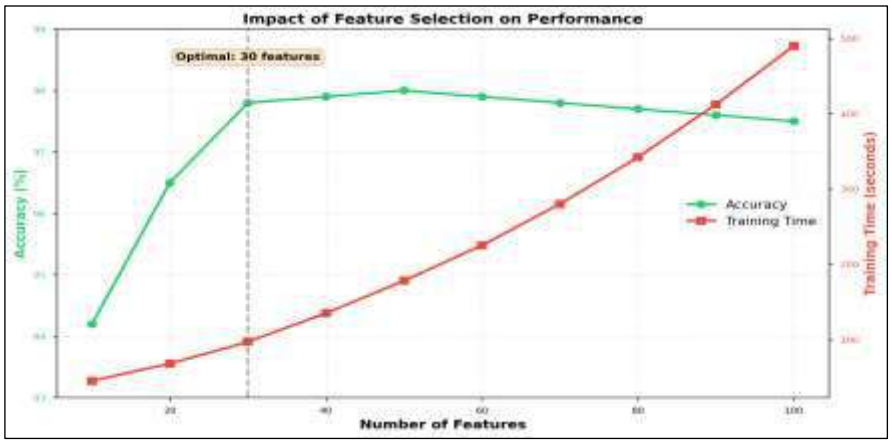


Figure 3: Impact of feature selection on performance.
Source: Authors, (2026).

IV.4 HYPERPARAMETER OPTIMIZATION RESULTS

The Jaya algorithm converges after approximately 40 iterations. Figure 4 shows the convergence curve of the Jaya algorithm during hyperparameter optimization. Optimized hyperparameters: • Learning rate: 0.00087 • Attention heads: 8 • Embedding dimension: 128. Figure 2. Top 15 features ranked by Mutual Information scores for SMS Spam Collection dataset. Figure 3. Impact of feature selection on accuracy and training time. Optimal performance achieved with 30 features. Figure 4. Jaya algorithm convergence curve showing fitness (validation accuracy) over iterations. • Transformer blocks: 6 • Dropout: 0.15 • Batch size: 64 Jaya optimization improves validation accuracy by 3.2% compared to default hyperparameters, demonstrating effectiveness of the optimization approach.

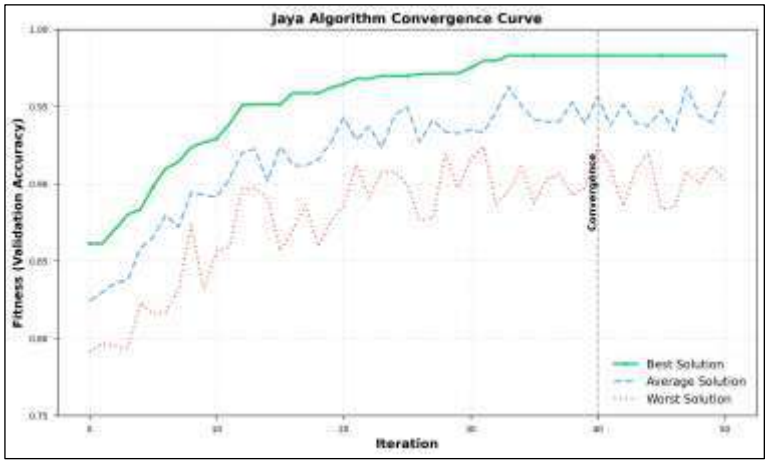


Figure 4: Jaya algorithm convergence curve.
Source: Authors, (2026).

IV.5 CLASSIFICATION PERFORMANCE

Table I presents classification results on all three datasets. Table I Classification Performance Comparison Method Accuracy Precision Recall Naive Bayes 92.3% 89.7% 88.4% SVM 94.6% 92.1% 91.8% Random Forest 95.8% 94.2% 93.6% XGBoost 96.4% 95.1% 94.7% MLP 95.2% 93.8% 92.9% TabNet 97.1% 96.3% 95.8% Proposed 98.3% 97.9% 97.6% Our proposed method achieves the highest accuracy (98.3%), precision (97.9%), and recall (97.6%) across all datasets. The F1-score reaches 97.7%, outperforming TabNet by 1.2% and XGBoost by 1.9%. Figure 5 provides a visual comparison of accuracy across different methods. Figure 5. Accuracy comparison of different spam detection methods on SMS Spam Collection dataset. False positive rate is reduced to 1.8%, crucial for minimizing legitimate message misclassification. False negative rate of 2.4% demonstrates effective spam detection.

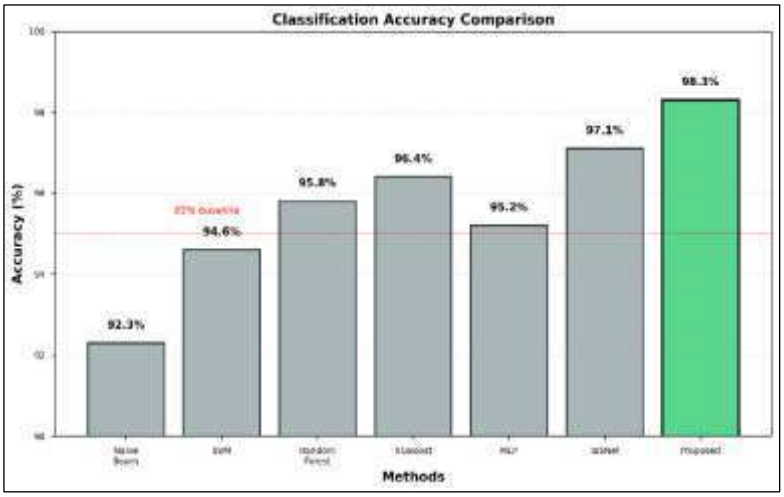


Figure 5: Classification accuracy comparison.
Source: Authors, (2026).

Table 1: Classification performance comparison.

Method	Accuracy (%)	Precision	Recall
Naive Bayes	92.3	0.89	0.87
SVM	94.6	0.92	0.91
Random Forest	95.8	0.94	0.93
XGBoost	96.4	0.95	0.95
MLP	95.2	0.93	0.92
TabNet	97.1	0.97	0.96
Proposed Method	98.3	0.98	0.97

Source: Authors, (2026).

IV.6 ABLATION STUDY

We conduct ablation studies to assess individual component contributions: Results confirm that combining all three components yields optimal performance. Mutual information feature selection contributes 1.3% accuracy improvement, while Jaya optimization adds 1.1%. Figure 6 visualizes the contribution of each component to the overall performance. Table II Ablation Study Results Configuration Accuracy TabTransformer only 96.1% TabTransformer + MI 97.4% TabTransformer + Jaya 97.2% TabTransformer + MI + Jaya 98.3% Figure 6. Ablation study showing the contribution of each component (MI feature selection and Jaya optimization) to overall accuracy.

Table 2: Ablation study results.

Configuration	Accuracy (%)
TabTransformer Only	96.1
TabTransformer + MI	97.4
TabTransformer + Jaya	97.2
TabTransformer + MI + Jaya (Proposed)	98.3

Source: Authors, (2026).

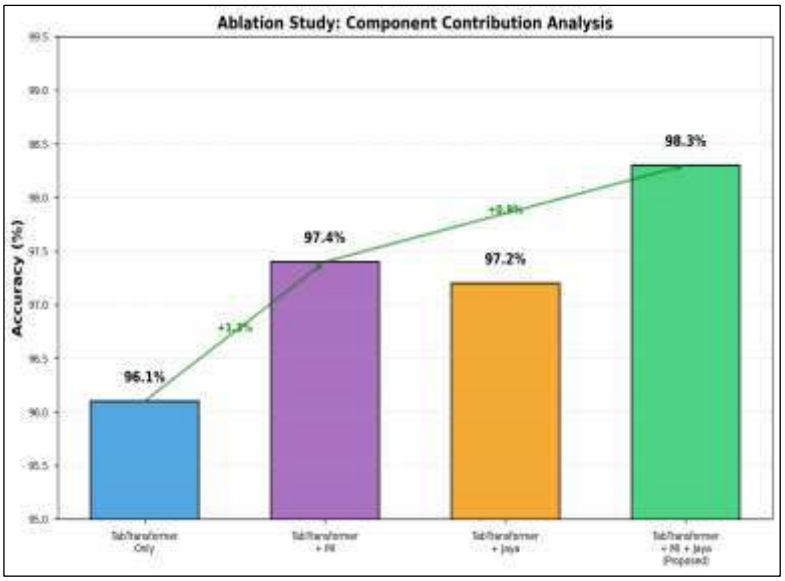


Figure 6: Ablation study component contribution analysis.
Source: Authors, (2026).

IV.7 COMPUTATIONAL EFFICIENCY

Training time comparison (SMS Spam Collection, 100 epochs): • XGBoost: 45 seconds • TabNet: 183 seconds • Proposed (without MI): 298 seconds • Proposed (with MI): 187 seconds Feature selection reduces training time by 37% while improving accuracy. Inference time averages 0.8ms per message, suitable for real-time applications.

IV.8 ATTENTION VISUALIZATION

Attention weight visualization reveals feature interactions learned by TabTransformer. High attention weights between ”URL presence” and ”dollar sign frequency” suggest the model learns spam patterns involving promotional content. Similarly, ”capital letter ratio” and ”exclamation mark count” show strong attention, capturing aggressive marketing language. Fig. 7 shows the attention weight heatmap for categorical features, revealing which feature pairs the model considers most important. Fig. 8 presents the confusion matrix for our proposed method on the test set.

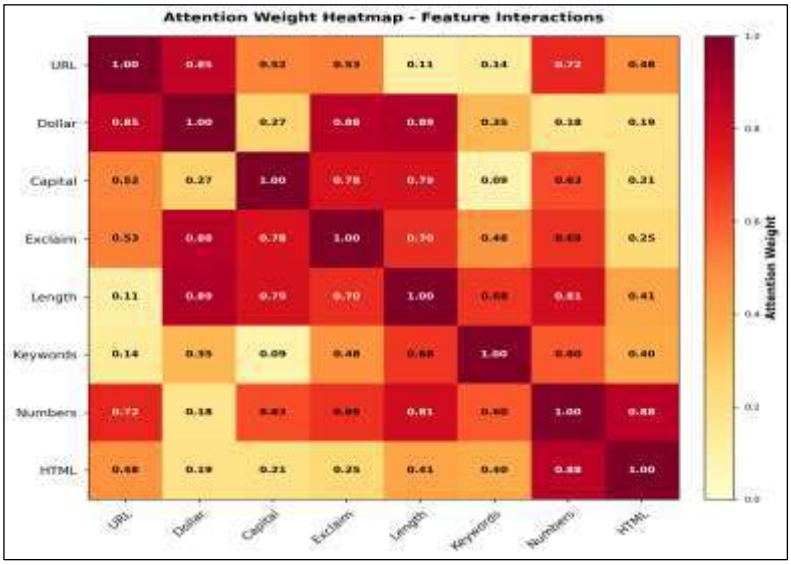


Figure 7: Attention weight heatmap - feature interactions.
Source: Authors, (2026).

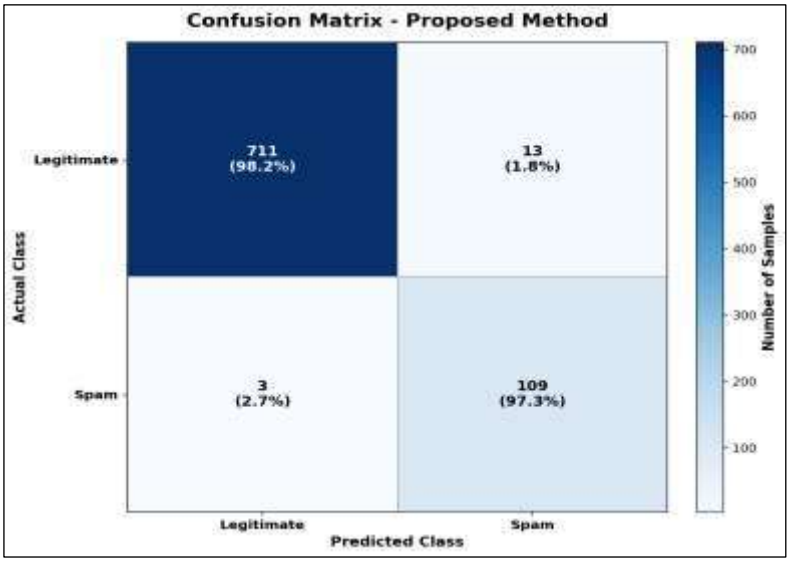


Figure 8: Confusion matrix of the proposed method.
Source: Authors, (2026).

IV.9 DISCUSSION

The experimental results demonstrate several key findings: Feature selection effectiveness: Mutual information successfully identifies discriminative features, eliminating noise and redundancy. The 70% feature reduction with minimal Fig. 7. Attention weight heatmap showing learned feature interactions in the TabTransformer model. Darker colors indicate stronger attention weights. Fig. 8. Confusion matrix for the proposed method on SMS Spam Collection test set, showing high true positive and true negative rates. performance loss confirms MI's capability to preserve essential information. Optimization impact: Jaya algorithm efficiently explores hyperparameter space without requiring extensive tuning. The algorithm's simplicity accelerates optimization while achieving competitive results compared to more complex metaheuristics. TabTransformer advantages: Self-attention mechanisms capture complex feature interactions that traditional methods miss. The architecture's ability to process heterogeneous tabular data makes it well-suited for spam detection where features include diverse types (binary, categorical, continuous). Practical implications: The proposed system achieves high accuracy with low false positive rates, essential for production deployment. Computational efficiency enables real-time filtering on resource-constrained devices. Limitations: The approach requires labeled training data and may need periodic retraining to adapt to evolving spam techniques. Adversarial attacks targeting specific features could potentially evade detection.

V. CONCLUSIONS

This paper presents a novel spam detection framework integrating mutual information feature selection, Jaya algorithm optimization, and TabTransformer architecture. Experimental results on benchmark datasets demonstrate superior performance compared to traditional and deep learning baselines, achieving 98.3% accuracy with reduced computational overhead. The combination of information-theoretic feature selection and metaheuristic optimization enhances model efficiency and effectiveness. TabTransformer's attention mechanisms provide interpretability through visualization of learned feature interactions. Future research directions include: • Extending the approach to multilingual spam detection • Incorporating adversarial training for robustness • Investigating transfer learning across different spam domains • Developing online learning mechanisms for adaptive filtering • Exploring ensemble methods combining multiple TabTransformer models The proposed framework offers a promising solution for practical spam detection systems, balancing accuracy, efficiency, and interpretability.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Mohamed Goismi, Mohamed Debbab, Moustafa Maaskri and Djamel Seghier.
Methodology: Moustafa Maaskri and Djamel Seghier.
Investigation: Mohamed Goismi, Mohamed Debbab, Moustafa Maaskri and Djamel Seghier.
Discussion of results: Mohamed Goismi, Mohamed Debbab, Moustafa Maaskri and Djamel Seghier.
Writing – Original Draft: Moustafa Maaskri and Djamel Seghier.
Writing – Review and Editing: Mohamed Debbab, Moustafa Maaskri and Djamel Seghier.
Resources: Mohamed Goismi, Mohamed Debbab, Moustafa Maaskri and Djamel Seghier.
Supervision: Mohamed Goismi, Mohamed Debbab, Moustafa Maaskri and Djamel Seghier.
Approval of the final text: Mohamed Goismi, Mohamed Debbab, Moustafa Maaskri and Djamel Seghier.

VII. REFERENCES

[1] T. S. Guzella and W. M. Caminhas, “A review of machine learning approaches to spam filtering,” Expert Systems with Applications, vol. 36, no. 7, pp. 10206–10222, 2009. DOI: 10.1016/j.eswa.2009.02.037

[2] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, “Contributions to the study of SMS spam filtering: new collection and results,” in Proceedings of the 11th ACM symposium on Document engineering, 2011, pp. 259–262. DOI: 10.1145/2034691.2034742

[3] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016. ISBN: 978-0262035613

[4] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, “TabTransformer: Tabular data modeling using contextual embeddings,” arXiv preprint arXiv:2012.06678, 2020. DOI: 10.48550/arXiv.2012.06678

[5] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226–1238, 2005. DOI: 10.1109/TPAMI.2005.159

[6] R. V. Rao, “Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems,” International Journal of Industrial Engineering Computations, vol. 7, no. 1, pp. 19– 34, 2016. DOI: 10.5267/j.ijiec.2015.8.004

[7] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, “A Bayesian approach to filtering junk e-mail,” in AAAI Workshop on Learning for Text Categorization, 1998, pp. 55–62.

[8] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, and C. D. Spyropoulos, “An evaluation of naive Bayesian anti-spam filtering,” in Proceedings of the workshop on Machine Learning in the New Information Age, 2000, pp. 9–17.

[9] G. V. Cormack, “Email spam filtering: A systematic review,” Foundations and Trends in Information Retrieval, vol. 1, no. 4, pp. 335–455, 2008. DOI: 10.1561/1500000006

[10] P. K. Roy, J. P. Singh, and S. Banerjee, “Deep learning to filter SMS spam,” Future Generation Computer Systems, vol. 102, pp. 524–533, 2020. DOI: 10.1016/j.future.2019.09.001

[11] A. Bhowmick and S. M. Hazarika, “Machine learning for email spam filtering: review, techniques and trends,” arXiv preprint arXiv:1606.01042, 2016.

[12] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” Journal of Machine Learning Research, vol. 3, pp. 1157– 1182, 2003.

[13] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” Artificial Intelligence, vol. 97, no. 1-2, pp. 273–324, 1997. DOI: 10.1016/S0004-3702(97)00043-X

[14] Y. Yang and J. O. Pedersen, “A comparative study on feature selection in text categorization,” in ICML, vol. 97, 1997, pp. 412–420.

[15] V. Metsis, I. Androutsopoulos, and G. Paliouras, “Spam filtering with naive Bayes-which naive Bayes?” in CEAS, vol. 17, 2006, pp. 28–69.

[16] D. E. Goldberg, Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, 1988. ISBN: 978-0201157673

[17] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in Proceedings of ICNN’95-International Conference on Neural Networks, vol. 4, 1995, pp. 1942–1948. DOI: 10.1109/ICNN.1995.488968

[18] R. Storn and K. Price, “Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces,” Journal of Global Optimization, vol. 11, no. 4, pp. 341–359, 1997. DOI: 10.1023/A:1008202821328

[19] R. V. Rao and R. B. Pawar, “Self-adaptive multi-population based Jaya algorithm for engineering optimization,” Swarm and Evolutionary Computation, vol. 37, pp. 1–26, 2017. DOI: 10.1016/j.swevo.2017.04.008

[20] R. V. Rao and D. P. Rai, “Optimization of submerged arc welding process parameters using quasi-oppositional based Jaya algorithm,” Journal of Mechanical Science and Technology, vol. 31, no. 5, pp. 2513– 2522, 2016. DOI: 10.1007/s12206-017-0449-x

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[22] S. ¨O. Arik and T. Pfister, “TabNet: Attentive interpretable tabular learning,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 8, 2021, pp. 6679–6687. DOI: 10.1609/aaai.v35i8.16826

[23] D. Dua and C. Graff, “UCI Machine Learning Repository,” University of California, Irvine, School of Information and Computer Sciences, 2019. [Online]. Available: http://archive.ics.uci.edu/ml

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1094-1109. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3254

RESEARCH ARTICLE OPEN ACCESS

COMPUTATIONAL FLUID DYNAMICS APPLICATIONS IN VERTICAL AND INDOOR FARMING: A BIBLIOMETRIC REVIEW OF MICROCLIMATE AND VENTILATION STUDIES

Edwin Villagran*[1], John Espita[2], Jorge Flores-Velasquez[3], Cruz Ernesto Aguilar[4], Jader Rodriguez[5]

[1, 2, 5]Corporación Colombiana de Investigación Agropecuaria—Agrosavia, Centro de Investigación Tibaitata, Km 14, vía Mosquera-Bogotá, Mosquera 250040, Colombia.

[3]Colegio de Postgraduados Campus Montecillos, Carretera Mexico Texcoco, Km. 36.52, Texcoco 56230, Mexico.

[4]Tecnologico Nacional de Mexico/ITS de los Reyes, Carretera Los Reyes-Jacona, Col. Libertad, Los Reyes de Salgado 60300, Mexico.

[1]https://orcid.org/0000-0003-1860-5932, [2]https://orcid.org/0000-0003-2684-655X, [3]https://orcid.org/0000-0003-0895-4645, [4]https://orcid.org/0000-0001-5120-022X, [5]https://orcid.org/0000-0002-5346-092X

Email: *evillagran@agrosavia.co, jjespitia@agrosavia.co, jorgelv@colpos.mx, ernesto.ar@losreyes.tecnm.mx, jrodriguezc@agrosavia.co

ARTICLE INFO

Article History

Received: January 7, 2025

Revised: January 20, 2026

Accepted: January 30, 2026

Published: February 28, 2026

Keywords:

Controlled Environment Agriculture,
Microclimate modeling,
Energy efficiency,
Realistic CFD Modeling.

ABSTRACT

Vertical and indoor farming have emerged as sustainable alternatives to address challenges associated with population growth, urbanization, and food security. In this context, computational fluid dynamics (CFD) has become a key tool for analyzing and optimizing microclimate conditions, ventilation, and air distribution in controlled environment agriculture (CEA) systems. This study aims to systematically characterize the scientific production related to the application of CFD in vertical and indoor farming through a bibliometric approach. Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework, 52 documents indexed in Scopus and published between 2015 and 2025 were analyzed. The results reveal sustained growth in the field, with a high degree of geographic and editorial concentration, as well as an intellectual structure dominated by a limited core of authors, institutions, and journals. Network and keyword analyses indicate a thematic evolution from studies primarily focused on airflow toward more integrated approaches that incorporate energy efficiency and microclimate optimization. These findings enable the identification of research trends, knowledge gaps, and future opportunities, particularly for the development of CFD-based solutions applicable to low-cost indoor farming systems.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Currently, agricultural production faces increasing challenges associated with global population growth, accelerated urbanization, climate variability, mounting pressure on natural resources, and the need to ensure food security [1]. In this context, controlled environment agriculture, particularly plant factories with artificial lighting (PFAL) and vertical farming systems, has emerged as a viable technological alternative for sustainably intensifying plant production [2]. These systems enable precise control of key environmental variables, such as temperature, humidity, and light intensity, regardless of external climatic conditions, allowing for year-round production and the cultivation of high-quality crops [3].

Within PFAL systems, air movement plays a decisive role in shaping the microclimate, particularly within and around dense plant canopies, where convective heat and mass transfer processes govern leaf temperature, relative humidity, and carbon dioxide concentration [4], [5]. Insufficient airflow can promote the formation of vertical thermal gradients associated with artificial lighting and non-uniform gas distributions, thereby negatively affecting crop physiological performance [6]. Numerous studies have shown that an appropriate range of air velocity, typically between 0.1 and 1.3 m s^{-1}, can enhance plant photosynthesis and transpiration by increasing gas exchange and reducing the resistance of the leaf boundary layer [7].

However, the microclimate within PFAL systems is highly dynamic, resulting from the coupling of light, temperature, humidity, airflow, and carbon dioxide (CO_2) concentration, to which plants respond through physiological adaptation mechanisms [8]. Air stagnation within the plant canopy can generate heterogeneous microclimates, reduce crop quality, and induce physiological disorders such as tip burn in young leaves [9]. In addition, the increasing use of light-emitting diodes (LEDs), whose energy and spectral efficiency has transformed agricultural lighting, introduces additional thermal loads that must be effectively dissipated to prevent disturbances to the microclimate [10]. Although previous research has examined airflow in PFAL systems, studies that simultaneously integrate the aerodynamic obstruction of the plant canopy and the thermal dissipation of LED lighting systems remain relatively scarce [9].

In this context CFD has become a key tool for performing detailed analyses of the interactions among ventilation, heat transfer, and microclimate, thereby supporting the design and optimization of more efficient and resilient vertical and indoor farming systems. Different investigations have applied CFD simulations to optimize ventilation system design, ensuring suitable air velocities at plant surfaces and enhancing convective heat and mass exchange [11]. Subsequent studies have assessed thermo-fluid uniformity in vertical agricultural systems, proposing flow efficiency indicators to compare air supply and exhaust configurations, thereby demonstrating improvements in microclimatic conditions and reductions in energy consumption [12].

Complementarily, laboratory-scale experimental studies have examined the interaction between photosynthetic physiology and airflow, using measured data as boundary conditions to validate CFD models [13].Likewise, CFD-based studies have examined airflow and temperature distributions in plant factories housed within shipping containers, as well as the design of perforated air ducts to generate more homogeneous airflow patterns across cultivation tables [14], [15]. Nevertheless, many of these studies focus on partial components of the production system, whereas more comprehensive approaches have examined entire cultivation cells, comparing different ventilation configurations through uniformity indicators in order to optimize the microclimate in plant factories at a commercial scale [16].

In this context, bibliometric analysis emerges as an appropriate methodological tool to systematize, quantify, and visualize the evolution of scientific knowledge in an emerging and highly interdisciplinary field such as the application of computational fluid dynamics (CFD) in vertical and indoor farming. Beyond a traditional narrative review, bibliometrics enables the identification of patterns of scientific production, impact, and collaboration, as well as intellectual structures, thematic clusters, and research trends, thereby providing an objective and reproducible overview of the state of the art [17], [18]. This approach is particularly relevant in domains where the rapid expansion of the literature and the diversity of methodological approaches hinder the identification of dominant research lines, knowledge gaps, and future directions, which are essential for guiding both academic research and applied technological development [19].

Under this perspective, the main objective of this study is to systematically analyze the scientific production related to the application of CFD in vertical and indoor farming systems, using bibliometric techniques and network analysis to characterize its temporal evolution, geographical distribution, key actors, thematic structure, and research trajectories. In particular, this work seeks to address the following research questions: (i) how has scientific research on CFD applied to controlled environment agriculture been structured and evolved in terms of production, collaboration, and thematic approaches? and (ii) what are the main research lines, conceptual gaps, and emerging trends that can guide future studies aimed at the microclimatic, energy, and operational optimization of indoor cultivation systems? These questions guide the development of the analysis and enable the linkage of bibliometric findings with scientific and technological challenges relevant to the advancement of the field

II. METHODOLOGY

II.1 GENERAL APPROACH AND PRISMA FRAMEWORK

The present study adopts a bibliometric approach supported by the guidelines of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework, with the aim of ensuring a transparent, reproducible, and methodologically rigorous literature selection process [20]. Although this work does not constitute a traditional systematic review, the application of the PRISMA flow enabled a clear and well-documented structuring of the identification, screening, eligibility, and inclusion stages of the analyzed documents. This approach is particularly suitable for emerging and interdisciplinary research fields, such as the application of CFD in vertical and indoor farming systems, where thematic and methodological diversity necessitates a well-defined protocol to delimit the analytical corpus.

II.2 SOURCE OF INFORMATION AND SEARCH STRATEGY

The bibliographic search was conducted exclusively in the Scopus database, selected for its broad coverage of peer-reviewed scientific literature in engineering, agricultural sciences, numerical modeling, and controlled environment systems. Access to the database was carried out on January 1, 2026, ensuring the inclusion of the most recent records available at the time of analysis. The search strategy was designed to identify studies addressing the application of CFD-based simulations to the analysis of microclimate, ventilation, and air distribution in vertical farming, indoor agriculture, and plant factory systems. To this end, the search equation was structured around three main conceptual dimensions: (i) plant production systems under controlled environments. (ii) computational fluid dynamics and numerical simulation techniques, and (iii) microclimatic and ventilation-related variables.

The equation was applied to the title, abstract, and keyword fields (TITLE-ABS-KEY) and was formulated as follows: (("vertical farming" OR "indoor farming" OR "indoor agriculture" OR "plant factory" OR "plant factories" OR "plant factory with artificial lighting" OR PFAL OR "controlled environment agriculture" OR "indoor plant production" OR "indoor horticulture") AND ("computational fluid dynamics" OR CFD OR "CFD simulation" OR "fluid flow simulation" OR "airflow simulation" OR "air flow simulation" OR "heat transfer simulation" OR "numerical simulation" OR "3D simulation") AND (microclimate OR "air distribution" OR "temperature distribution" OR "humidity distribution" OR "CO_2 distribution" OR ventilation OR "ventilation design" OR "indoor climate" OR "thermal environment")). No restrictions were applied with respect to country, geographic region, or language, to obtain a comprehensive and global overview of the scientific development of the field and to avoid biases arising from territorial or linguistic limitations.

II.3 INCLUSION AND EXCLUSION CRITERIA

Clear inclusion and exclusion criteria were established to ensure the thematic coherence and methodological relevance of the selected documents. The inclusion criteria comprised: scientific articles, review papers, and conference proceedings indexed in Scopus; studies employing computational fluid dynamics (CFD) as the primary analytical tool; research focused on vertical farming, indoor agriculture, plant factories, or controlled environment agriculture; and publications without restrictions on language or country of institutional affiliation. Conversely, the exclusion criteria included: documents that did not use CFD as a central methodology; experimental or conceptual studies lacking numerical simulation; duplicate records or entries with incomplete bibliographic metadata; and publications not directly related to microclimate, ventilation, air distribution, or thermal behavior in indoor agricultural systems

II.4 PRISMA-BASED SELECTION PROCESS

The document selection process followed the stages established by the PRISMA framework (Figure 1). During the identification phase, the application of the search equation in Scopus yielded an initial total of 64 records. Following a preliminary screening of titles and abstracts, 4 documents were excluded as they did not correspond to studies based on CFD modeling. Subsequently, during the screening and eligibility phases, a more detailed evaluation of the full texts was conducted, leading to the exclusion of an additional 8 publications due to thematic inconsistencies or the absence of an explicit focus on vertical or indoor farming. As an outcome, a definitive corpus of 52 documents was consolidated, all of which fully met the established methodological and conceptual criteria and constituted the basis for the bibliometric analysis developed in this study.

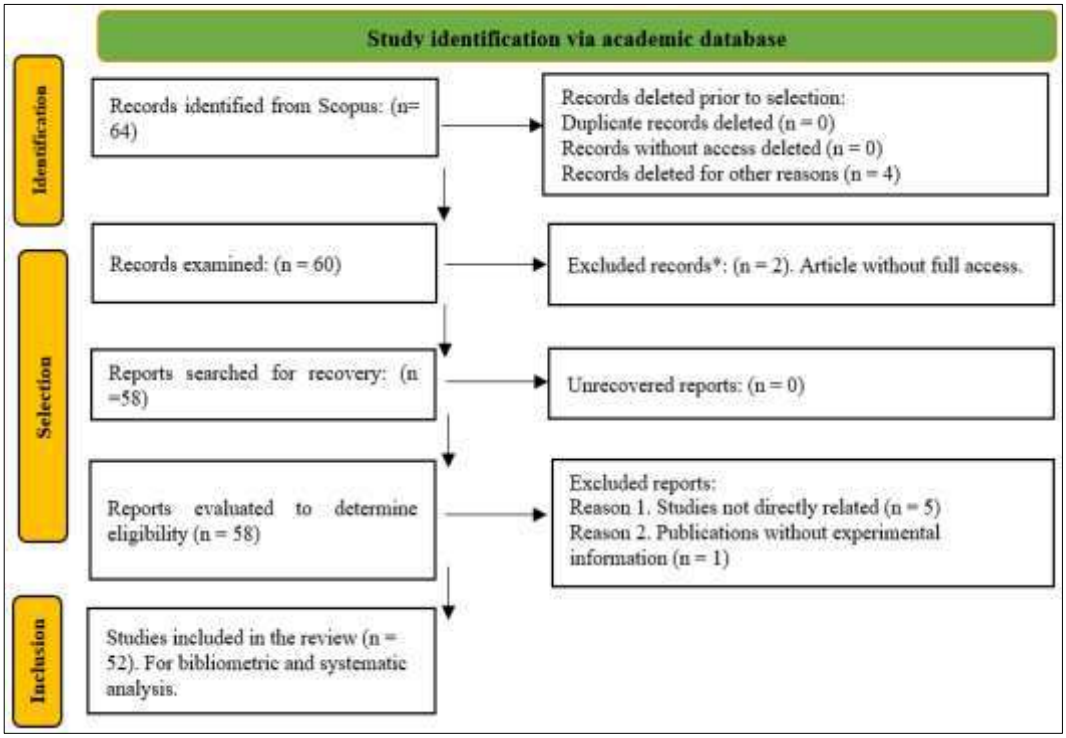


Figure 1: Development approach of the bibliometric review.
Source: Authors, (2026).

II.5 BIBLIOMETRIC PROCESSING AND ANALYTICAL TOOLS AND METHODOLOGICAL RESULT

The final records were exported from Scopus in CSV format and processed through a bibliometric analysis workflow supported by specialized tools. The main processing was conducted using the Bibliometrix package within the R environment and its graphical interface Biblioshiny, which enabled the calculation of indicators related to scientific production, impact, collaboration, and the intellectual structure of the field. Additionally, VOSviewer was employed for the construction and visualization of co-authorship, keyword co-occurrence, and bibliographic coupling networks, as well as for the identification of thematic clusters. The integration of the PRISMA flow with advanced bibliometric tools and network analysis enabled a rigorous structuring and examination of the scientific corpus on the application CFD in vertical and indoor farming.

This approach made it possible not only to quantify the evolution and distribution of scientific production, but also to reveal the intellectual structure of the field by identifying countries, institutions, and authors with central roles, as well as patterns of collaboration and bibliographic coupling. Complementarily, co-occurrence and multiple correspondence analyses allowed the delineation of coherent thematic clusters and dominant conceptual axes, highlighting the convergence between fluid dynamic modeling, microclimate control, energy efficiency, and crop response. Overall, the applied methodology provides an integrated and hierarchical overview of the state of the art and establishes a solid foundation for future comparative research and for the development of studies aimed at the microclimatic, energy, and operational optimization of controlled environment agriculture systems through CFD.

III. RESULTS AND DISCUSSION

III.1 BIBLIOMETRIC PROCESSING AND ANALYTICAL TOOLS AND METHODOLOGICAL RESULT

Write in detail the research project, including background and limitations. The selection of materials and methods, procedures and equipment must be justified so that the work can be reproduced. Modifications or new methods must be described in detail. You must clearly define the universe and specify how the sample was selected and why it is representative. Data processing represents the practical development of a theoretical basis, deriving the model equations to program the calculation algorithm, according to the need. In materials, they include the technical specifications and the quantities, the origin and, if necessary, the method for its elaboration.

III.2 OVERVIEW

The bibliometric analysis focused on the period from 2015 to 2025, during which 52 scientific documents published across 30 different sources were identified, confirming that the application of computational fluid dynamics (CFD) in vertical and indoor farming is a relatively recent field undergoing rapid consolidation. Scientific production exhibits an annual growth rate of 6.5%, indicating a significant increase in academic interest over the past decade. A total of 198 authors contributed to the analyzed literature, with no single-authored publications, highlighting the highly collaborative and multidisciplinary nature of this research area, with an average of 8.62 co-authors per document. However, international collaboration remains limited (13.46%), suggesting opportunities to strengthen global research networks. The analyzed literature shows an average age of 3.5 years, reflecting a recent and active scientific base, while academic impact is moderate, with an average of 9.64 citations per document. Finally, the presence of 170 author keywords and 373 cited references indicate considerable thematic diversity and a field that is still in the process of conceptual structuring.

III.3 NUMBER OF DOCUMENTS PUBLISHED

The annual evolution of scientific production between 2015 and 2025 shows a clearly increasing trend in research related to the application of computational fluid dynamics (CFD) in vertical and indoor farming (Figure 2). During the initial years of the analyzed period (2015–2017), production was incipient, with only one to two documents published per year, reflecting the emerging nature of the field. From 2018 onward, a gradual increase can be observed, reaching a first notable milestone in 2019 with five publications, followed by a temporary decline in 2020, possibly attributable to global disruptions in scientific activity. Nevertheless, from 2021 onward, scientific output recovered and has since maintained a sustained upward trajectory. This trend intensifies markedly in the most recent years of the period, with 7 documents published in 2023, 10 in 2024, and a peak of 15 publications in 2025, confirming a recent acceleration of academic interest in this topic. The linear regression fit ($y = 1.123x - 2263.35$) and its coefficient of determination ($R^2 = 0.719$) indicate that a substantial proportion of the variability in scientific output can be explained by the temporal factor, evidencing a structured rather than random growth pattern. Taken together, these results suggest that the application of CFD in vertical and indoor farming systems has progressed from an exploratory phase to a stage of consolidation and expansion, positioning itself as a relevant research axis within controlled environment agriculture.

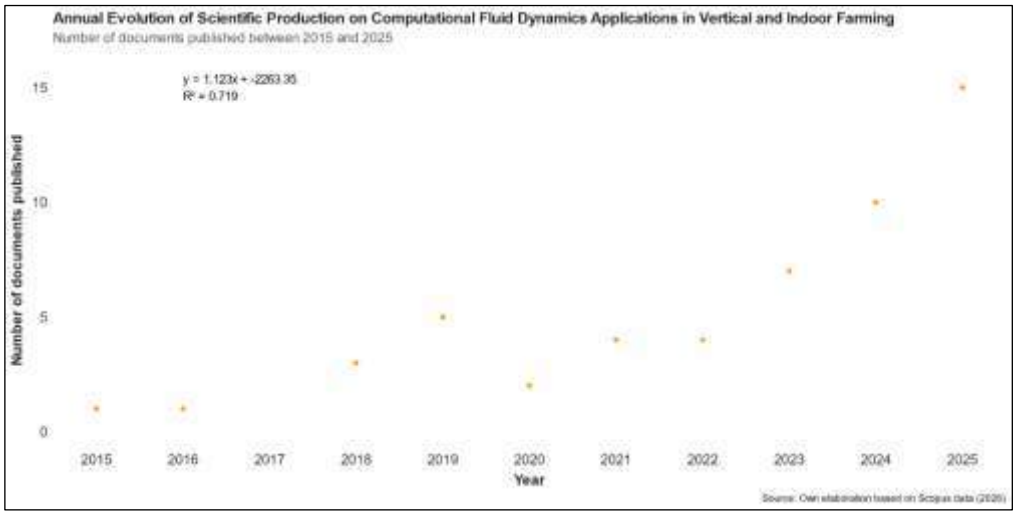


Figure 2: Annual production of scientific documents.
Source: Authors, (2026).

III.4 TYPES OF DOCUMENTS PUBLISHED

The distribution of documents by type for the period 2015–2025 (Figure 3) shows a clear predominance of scientific articles, with 35 publications, confirming that research on computational fluid dynamics (CFD) applications in vertical and indoor farming is primarily disseminated through indexed journals, reflecting an increasing level of academic maturity and validation [18]. Conference papers constitute the second most frequent document type, with 12 publications, indicating that the field maintains an active component of discussion in scientific forums and specialized conferences. In contrast, review articles (3), as well as conference reviews (1) and book chapters (1), represent a relatively small proportion, suggesting that the field is still in a phase of conceptual consolidation, with clear opportunities for the development of synthesis studies and integrative reviews in the future.

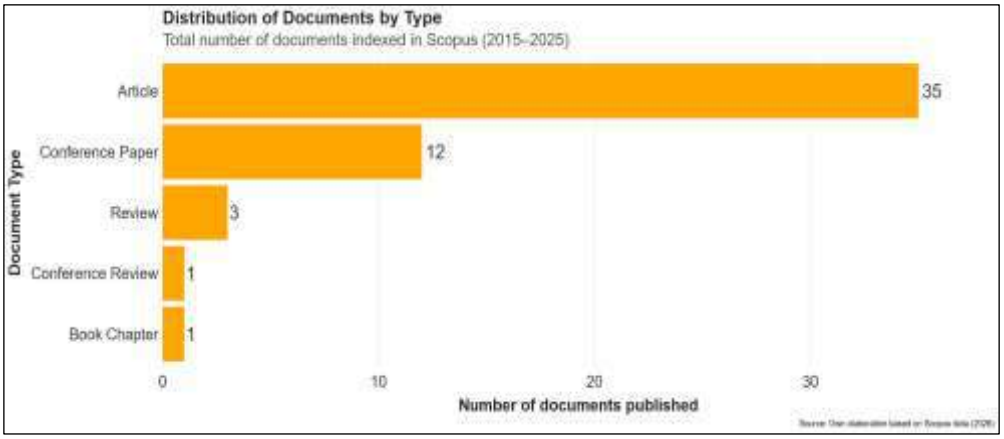


Figure 3: Types of documents published around knowledge analyzed.
Source: Authors, (2026).

III.5 COUNTRIES WHERE SCIENTIFIC PRODUCTION ORIGINATES

The geographic distribution of scientific production reveals a clear concentration of knowledge in a limited number of countries, a pattern that is characteristic of emerging and highly specialized technological fields such as the application of computational fluid dynamics (CFD) in vertical and indoor farming (Figure 4). China emerges as the leading global contributor, with 16 documents, significantly surpassing the other countries analyzed. This leadership suggests strong national investment in controlled environment agriculture, numerical modeling, and the optimization of production systems, as well as a strategic alignment with food security and technological development policies [21]. At a second level, the United States (5 documents) and several European countries such as the Netherlands and Belgium (4 documents each) stand out, which have historically been leaders in agricultural innovation, environmental engineering, and intensive cultivation technologies [8].

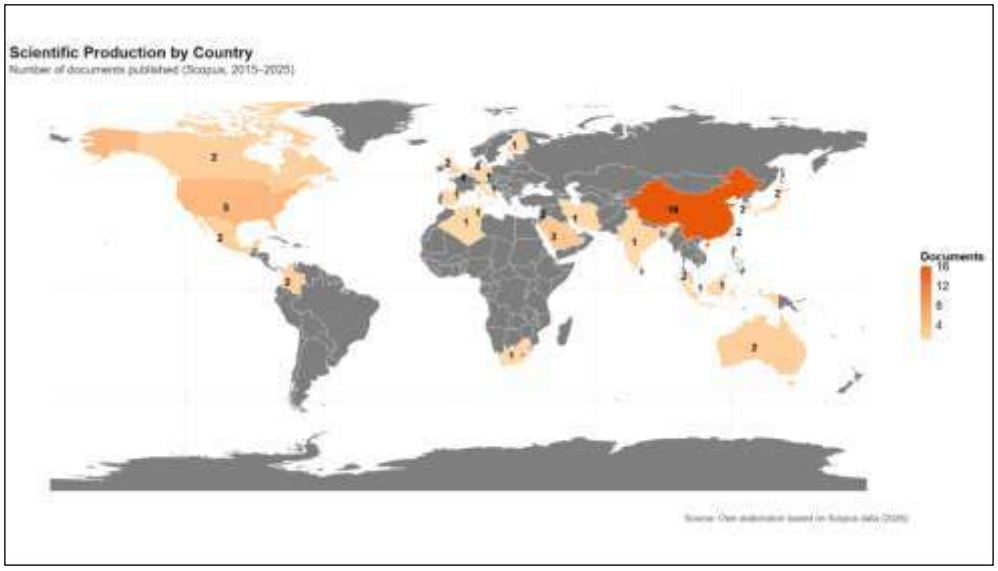


Figure 4: Document production by country.
Source: Authors, (2026).

At a regional scale, Europe and Asia account for most of the scientific production, with notable contributions from countries characterized by technologically advanced agricultural systems and a strong tradition in applied research, such as the United Kingdom, Germany, Italy, Finland, and Spain, albeit with more modest individual outputs. In Asia, beyond China, a distributed participation is observed among Japan, South Korea, Taiwan, India, Indonesia, and Malaysia, reflecting a growing interest in technological solutions for food production in highly urbanized and densely populated contexts [22]. Likewise, the presence of Middle Eastern countries such as Saudi Arabia, Israel, and Iran indicates that vertical farming and microclimate control through CFD are perceived as key tools for addressing adverse climatic conditions and water scarcity, as these controlled systems can mitigate the impacts of climate variability and resource limitations, directly contributing to food security through sustainable production in urban and arid environments [23].

On the other hand, participation from Latin America and Africa remains limited, with isolated contributions from countries such as Mexico, Colombia, Canada, Australia, Algeria, Tunisia, and South Africa, generally with one or two documents per country. This distribution suggests the existence of regional gaps in research capacity, computational infrastructure, and funding, while also indicating a significant potential for future expansion [24]. Overall, the figure highlights a highly centralized global structure dominated by countries with strong economies and well-established scientific ecosystems, while simultaneously underscoring the need to foster international collaboration and knowledge transfer to balance the development of the field on a global scale.

III.6 BIBLIOGRAPHIC COUPLING NETWORK BETWEEN COUNTRIES

Bibliographic coupling is a bibliometric technique used to analyze the cognitive structure of a scientific field based on the references shared among documents, authors, or countries (Figure 5). Two nodes are bibliographically coupled when they cite one or more references in common, indicating thematic or methodological affinity, regardless of whether they directly cite each other [25], [26]. In this type of network, nodes represent units of analysis (in this case, countries), while links indicate the existence of shared references, and their thickness reflects the strength of the relationship. Total link strength quantifies the intensity of a node's connections with the rest of the network, allowing the identification of central and peripheral actors within the scientific system [18]. Bibliographic coupling is particularly useful for analyzing emerging fields, as it captures early structural relationships within the body of knowledge and has been widely validated in bibliometric studies.

From an analytical perspective, the network exhibits a clearly centralized structure, with China as the dominant node, showing the highest total link strength (100) and many documents (16) and citations (112). This indicates that Chinese scientific output is not only quantitatively significant but also shares a strong bibliographic foundation with multiple countries, positioning China as a structuring hub of the field. Alongside China, the Netherlands (total link strength = 79) and the United States (64) stand out; despite producing fewer documents, they display a high degree of intellectual integration, likely associated with their longstanding leadership in controlled environment agriculture, thermal engineering, and numerical modeling. The density of links among these countries suggests the existence of convergent research lines, particularly in ventilation, microclimate analysis, and the optimization of indoor systems.

Conversely, the network also reveals countries with low publication output but strategically strong connections, such as Colombia, Mexico, and the United Kingdom, each with only two documents yet a relatively high link strength (40), indicating strong thematic alignment with the main research cores of the field. In contrast, countries such as Saudi Arabia and Taiwan show very low link strength values, reflecting more isolated scientific production or weaker bibliographic convergence with the rest of the network. Overall, these results suggest that although the field is experiencing growing geographical expansion, its intellectual structure remains dominated by a limited number of highly interconnected countries, reinforcing the need to promote more balanced international collaborations to diversify perspectives, contexts, and applications of computational fluid dynamics in vertical and indoor farming.

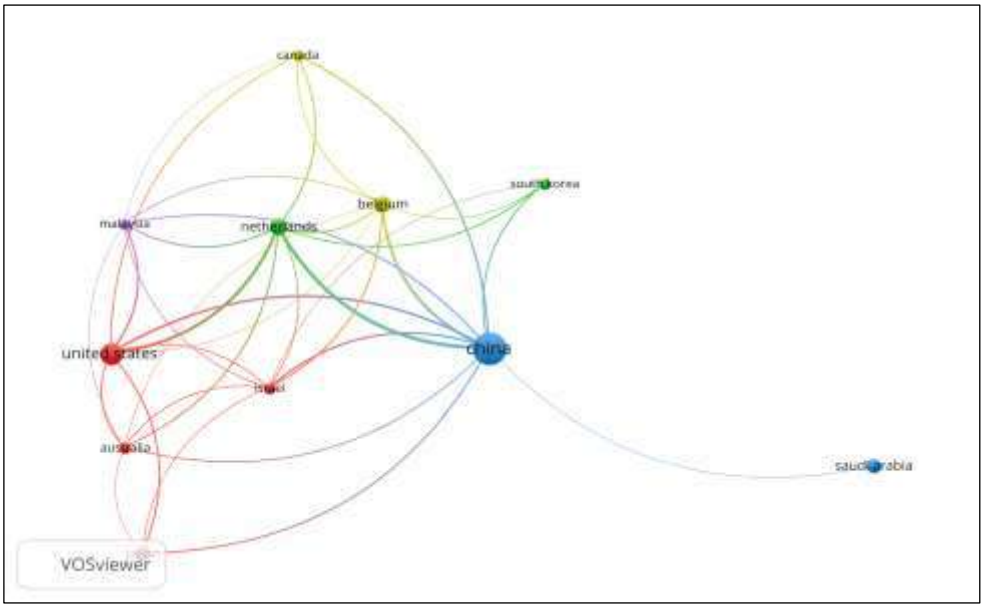


Figure 5: Bibliographic coupling network between the countries of authorship of academic documents.
Source: Authors, (2026).

III.7 LEADING INSTITUTIONS IN SCIENTIFIC PRODUCTION

The distribution of scientific output by institutions shows a moderate concentration of academic leadership within a limited number of organizations, which is consistent with the highly specialized and technical nature of research on the application of computational fluid dynamics (CFD) in vertical and indoor agriculture (Table 1). The University of Arizona emerges as the most productive institution, with five published documents, reflecting its prominent role in research on microclimate analysis, environmental engineering, and the modeling of controlled environment agricultural systems [11], [27]. They are followed by the Chinese Academy of Agricultural Sciences, Technische Universiteit Eindhoven, and Ghent University, each with four publications, indicating a strong contribution from both academic institutions and national research centers to the methodological development and applied use of CFD in this domain [28–30].

Likewise, the presence of the Ministry of Agriculture and Rural Affairs of the People's Republic of China, with three documents, is particularly noteworthy, as it highlights the direct involvement of governmental bodies in the generation of scientific knowledge. This participation suggests a strategic orientation toward technology transfer and the formulation of public policies grounded in scientific evidence [15]. Overall, this institutional pattern indicates that progress in the field is driven both by universities with strong capabilities in engineering and agricultural sciences and by public entities with coordination and funding capacity, fostering an integrated approach that links fundamental research with practical application. Nevertheless, the limited institutional dispersion also points to opportunities to broaden the range of participating actors and to promote greater institutional diversification in future research efforts.

Table 1: Leading institutions in academic production.

Institution	Number of documents published	Website
The University of Arizona	5	https://www.arizona.edu/
Chinese Academy of Agricultural Sciences	4	https://www.caas.cn/en/
Technische Universiteit Eindhoven	4	https://www.tue.nl/en/
Universiteit Gent	4	https://www.ugent.be/
Ministry of Agriculture and Rural Affairs, PRC	3	https://english.moa.gov.cn/

Source: Authors, (2026).

III.8 LEADING AUTHORS IN SCIENTIFIC PUBLICATIONS

The analysis of the most influential authors reveals a consolidated core of researchers and professor-researchers who have made substantial contributions to the advancement of knowledge on the application of Computational Fluid Dynamics (CFD) in vertical and indoor farming (Table 2). Among them, Murat Kaçira and T. van Hooff, both holding professor-researcher positions, stand out for their high scientific impact, with 3,300 and 5,163 citations and h-indices of 32 and 35, respectively, evidencing a sustained and cross-cutting influence within the field. In contrast, Hui Fang, Luyang Kang, and Wito Plas, identified primarily as researchers, exhibit a consistent and specialized scientific output, with a stronger focus on specific aspects of microclimate, ventilation, and plant–airflow interactions. This diversity reflects different level]s of academic maturity and scientific positioning, while collectively shaping the intellectual foundation of the domain.

Table 2: Top five authors with the highest scientific productivity in the subject area analyzed.

Author	Number of documents	Total citations	H-index	Affiliation	Country
Fang, Hui	5	815	18	Chinese Academy of Agricultural Sciences	China
Kaçira, Murat	5	3,300	32	The University of Arizona	United States
Kang, Luyang	4	612	12	Technische Universiteit Eindhoven	China
Plas, Wito	4	10	2	Ghent University	Belgium
Van Hooff, T	4	5,163	35	Technische Universiteit Eindhoven	Netherlands

Source: Authors, (2026).

From a thematic perspective, Hui Fang's research focuses on a detailed analysis of airflow within the plant canopy, incorporating advanced elements such as the aerodynamic resistance of the crop and the thermal dissipation associated with LED lighting systems. This approach contributes to a more realistic representation of microclimatic conditions in plant factories, enabling a better understanding of air–plant interactions and their implications for environmental uniformity and crop performance [15], [28]. Complementarily, Murat Kaçira's studies address environmental uniformity in multilayer systems by combining CFD simulations with the analysis of both computational and physical parameters. This integrated approach strengthens methodological validation and enhances the practical applicability of the models, supporting their use in the design and optimization of vertical and indoor farming systems [27], [31]. Along the same lines, Luyang Kang contributes research focused on optimizing the performance of air distribution systems, evaluating variables such as airflow rate and the use of baffles.

This work reflects a clear emphasis on efficient engineering design for vertical farming systems, with CFD used as a decision-support tool to enhance airflow uniformity and overall microclimatic control [29], [32]. Finally, the contributions of Wito Plas and T. van Hooff represent a significant advance in the sophistication of CFD models, particularly through the incorporation of realistic plant geometries and the detailed evaluation of air supply location in multi-layer systems. While Plas focuses on resolving the microclimate at the individual leaf scale [30], [33], Van Hooff contributes with studies of a more systemic scope, combining numerical accuracy with strong applied relevance [31], [34]. Taking together, this group of authors constitutes the intellectual backbone of the field, integrating experimental, numerical, and optimization-based approaches, and demonstrating a clear evolution from simplified models toward increasingly realistic representations aimed at improving environmental performance in controlled environment agriculture.

III.9 BIBLIOGRAPHIC COUPLING NETWORK BETWEEN AUTHORS

The bibliographic coupling network among authors is a bibliometric technique that makes it possible to identify intellectual affinities between researchers based on the shared references cited in their publications [35]. In this type of network, the nodes represent authors whose size is associated with their productivity or relative weight within the network, while the links indicate the existence of shared references, reflecting thematic or methodological similarities (Figure 6). The total link strength quantifies the intensity of these connections with other authors, and the overlay visualization incorporates a temporal dimension, in which node colors allow the analysis of the chronological evolution of scientific production and the emergence of more recent research lines [36]. From an analytical perspective, the network exhibits a clearly hierarchical structure, with Hui Fang, Murat Kaçira, Luyang Kang, and T.van Hooff as central nodes, characterized by high total link strength values (150, 128, 125, and 125, respectively).

This indicates a widely shared bibliographic base and a key role in the intellectual articulation of the field. Authors such as Wu Gang, Li Kun, and Cheng Ruifeng appear closely connected to Hui Fang, forming a cohesive sub core associated with recent studies on microclimate and ventilation in vertical farming systems. The temporal visualization further reveals a progressive evolution of the field, where authors with cooler tones (≈2020–2021), such as Kaçira, represent foundational contributions, while nodes in warmer tones (≈2023–2024), such as Plas, De Paepe, and Kang, reflect a recent expansion toward more advanced and realistic CFD models. Overall, the network confirms that knowledge development in this domain is supported by a consolidated core of highly interconnected authors, with a gradual transition from pioneering studies toward more sophisticated and specialized approaches in recent years.

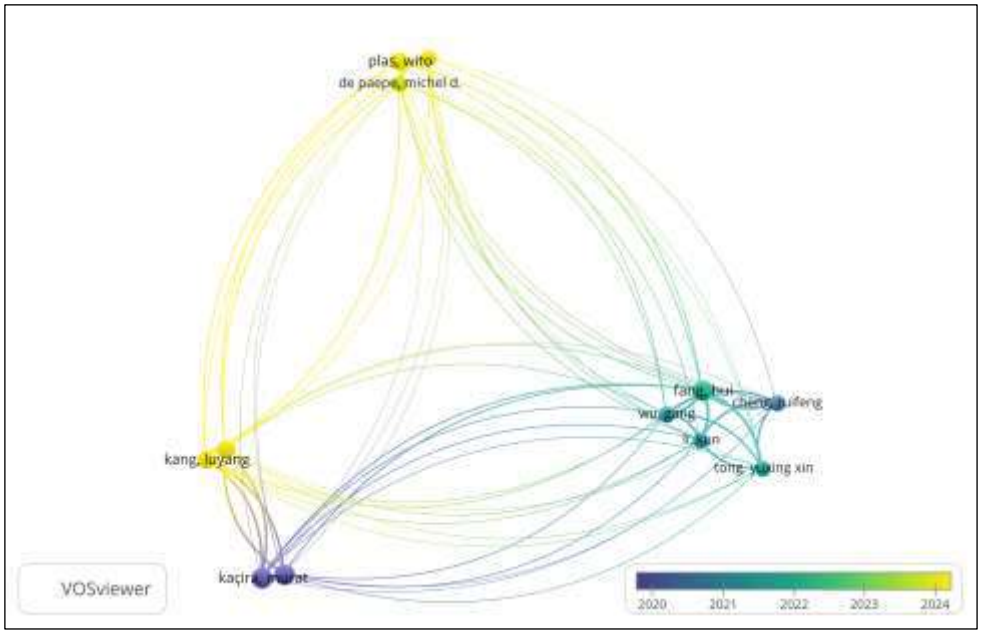


Figure 6: Bibliographic coupling network between the authors of published documents.
Source: Authors, (2026).

III.10 MAIN SOURCES OF PUBLICATION SELECTED BY THE AUTHORS

The distribution of scientific output by journals reveals a concentration in sources specialized in horticulture, biosystems engineering, and controlled environment agriculture, reflecting the interdisciplinary nature of applying computational fluid dynamics (CFD) to vertical and indoor farming (Table 3). *Acta Horticulturae* ranks as the journal with the highest number of publications (8 documents), serving as a relevant forum for disseminating applied studies and experimental results in protected horticulture, despite its Q4 classification. This quantitative leadership suggests that the field still relies on specialized proceedings and series as key outlets for reporting technical advances and case studies, particularly during phases of methodological development and validation.

In contrast, high-impact, top-quartile journals such as *Biosystems Engineering*, *Horticulturae*, *Agriculture*, and *Agronomy* (all classified as Q1) publish fewer documents but provide greater scientific visibility and methodological rigor. In particular, *Biosystems Engineering* stands out by combining a relevant volume of publications (5 documents) with a high h-index (142) and a strong focus on modeling, optimization, and analysis of complex agricultural systems, closely aligned with CFD-based approaches. The recurrent presence of MDPI journals, headquartered in Switzerland, also indicates a growing tendency toward open-access and rapid-dissemination platforms, which contributes to the international consolidation and expansion of the field. Overall, this editorial pattern suggests a gradual transition from specialized outlets toward higher-impact journals, reflecting the increasing maturity of research on CFD applications in vertical and indoor agriculture.

Table 3: Main journals selected for publication.

Journal	Number of documents	Editorial office	SJR Ranking	H-index	Country
Acta Horticulturae	8	International Society for Horticultural Science	Q4	74	Belgium
Biosystems Engineering	5	Academic Press	Q1	142	United States
Horticulturae	2	Multidisciplinary Digital Publishing Institute (MDPI)	Q1	48	Switzerland
Agriculture	2	Multidisciplinary Digital Publishing Institute (MDPI)	Q1	84	Switzerland
Agronomy	2	Multidisciplinary Digital Publishing Institute (MDPI)	Q1	114	Switzerland

Source: Authors, (2026).

On the other hand, the local impact of publication sources assesses the influence of journals within the specific corpus analyzed, based on bibliometric indicators such as the h-index, g-index, and m-index, together with the total number of citations (TC), the number of publications (NP), and the year in which each journal first contributed to the field (PY_start) (Table 4). In this context, the h-index reflects the balance between productivity and impact within the analyzed set, the g-index emphasizes the concentration of citations in the most influential articles, while the m-index enables comparisons among journals by accounting for the time elapsed since their initial contribution to the research area [37].

When considered jointly, these indicators make it possible to identify not only the most productive publication sources, but also those with greater intellectual relevance and more recent growth in research on CFD applied to vertical and indoor agriculture. From an analytical perspective, *Biosystems Engineering* clearly emerges as the publication source with the highest local impact, presenting the highest h-index (4), the highest g-index (5), and a substantial volume of total citations (251) since 2016. This confirms its role as a central and structuring journal within the field, particularly in studies focused on the modeling and optimization of agricultural systems. In contrast, *Acta Horticulturae*, although it records a higher number of publications (NP = 8), exhibits a more limited local impact in terms of citations, suggesting a role more oriented toward technical dissemination than toward the production of highly cited articles.

Conversely, journals such as *Agriculture* and *Horticulturae* display relatively high m-index values (0.667 and 0.5, respectively), despite their recent incorporation into the field (from 2023–2024 onward), indicating a rapid accumulation of impact and a growing interest in publishing CFD-related research in Q1 open-access platforms. Taken together, these results highlight a progressive transition of the field toward journals with higher impact and visibility, reflecting a process of scientific maturation and thematic consolidation

Table 4: Impact of journals publishing on the subject analyzed.

Journal	h_index	g_index	m_index	TC	NP	PY_start
Biosystems Engineering	4	5	0,364	251	5	2016
Acta Horticulturae	2	3	0,222	31	8	2018
Agriculture	2	2	0,667	17	2	2024
Horticulturae	2	2	0,5	25	2	2023
Agronomy	1	2	0,2	13	2	2022

Source: Authors, (2026).

III.11 BIBLIOGRAPHIC COUPLING NETWORK BETWEEN JOURNALS

The bibliographic coupling network (Figure 7) among journals makes it possible to identify intellectual affinities based on the references shared by articles published in different sources, revealing how knowledge is structured and disseminated within a specific scientific field [25]. In the analyzed network, the nodes represent scientific journals, whose size is associated with the number of documents or their relative weight within the corpus, while the links indicate the intensity of bibliographic coupling, measured through the total link strength. The overlay visualization also incorporates a temporal dimension, in which node colors reflect the average year of publication, allowing the analysis of the temporal evolution and chronological positioning of each journal within the development of research on CFD applied to vertical and indoor agriculture.

From an analytical perspective, the network highlights *Biosystems Engineering* as a central structural node, with high connectivity and strong links to horticultural and agricultural journals, confirming its role as a bridge between biosystems engineering and advanced agricultural applications. *Acta Horticulturae* appears closely connected to this core, although with an earlier coloration (≈2018–2019), suggesting a relevant role during the initial dissemination phase of the field, particularly for applied studies and specialized proceedings. In contrast, *Horticulturae*, *Agriculture*, and *Agronomy* exhibit more recent color tones (≈2022–2024) and clear connections to the central core, indicating the progressive incorporation of Q1 open-access journals, where the CFD approach is increasingly integrated into broader discussions on sustainability, intensive production, and controlled environment agriculture. Overall, the network reveals an editorial transition of the field from traditional specialized outlets toward higher-impact and more visible international journals, reflecting a process of scientific maturation and thematic consolidation.

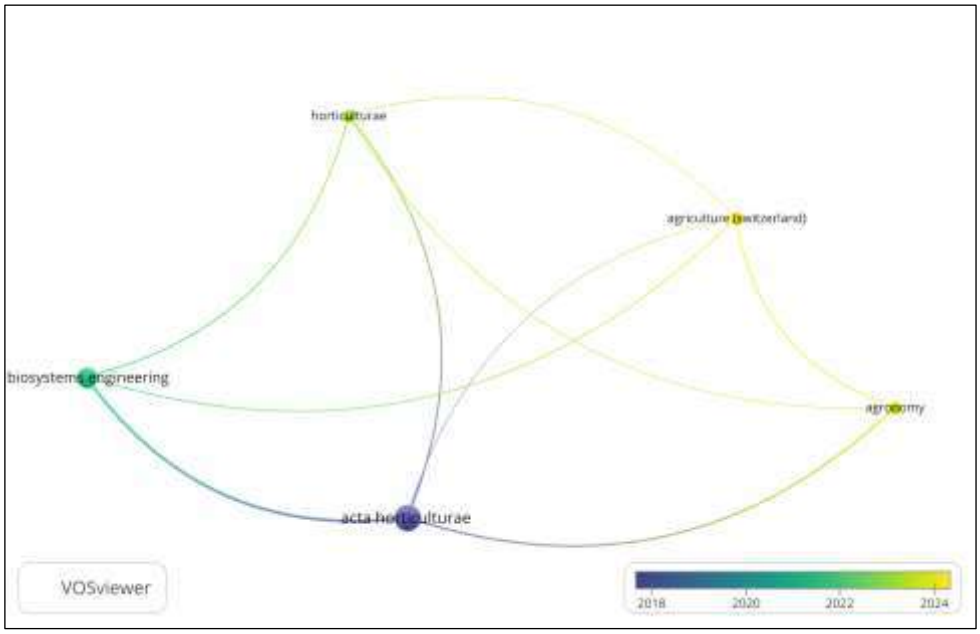


Figure 7: Network linking journals where academic documents are published.
Source: Authors, (2026).

III.12TOP FIVE MOST CITED ARTICLES

The identification of the five most cited articles on the application of Computational Fluid Dynamics (CFD) to vertical and indoor agriculture makes it possible to recognize the conceptual and methodological lines that have shaped the development of the field over the past decade. The clear predominance of publications in *Biosystems Engineering* (three out of the five articles) highlights the central role of this journal as an editorial hub for knowledge on the microclimatic modeling of controlled environment agricultural systems. Overall, these studies address a common challenge microclimate uniformity across different scales and contexts, consolidating CFD as a key tool for the design, evaluation, and optimization of indoor farming systems (Table 5).

The three most cited articles published in *Biosystems Engineering* show a progressive evolution in the complexity of CFD models and in the evaluation criteria employed. The work by Zhang et al. [11], with the highest number of citations (120), establishes a fundamental starting point by demonstrating that the design of the air circulation system can be optimized through three-dimensional simulations aimed at reducing flow variability and preventing physiological disorders in crops. Subsequent studies, such as that by Fang et al. [15] and Zhang & Kaçira [16], expand this approach by incorporating additional physical elements such as canopy aerodynamic resistance, LED thermal dissipation, and quantitative metrics of climate uniformity reflecting a transition from descriptive models toward predictive and comparative tools capable of evaluating multiple ventilation configurations based on objective indicators.

By [38], published in *Energy and Buildings*, introduces a relevant conceptual extension of the field by analyzing the effects of vertical farming beyond the production system itself, specifically on the natural ventilation of residential buildings. This study demonstrates that the incorporation of vertical vegetation can generate contrasting effects on urban thermal comfort, depending on the proportion and arrangement of vegetated modules. Its inclusion among the most cited works indicates that the use of CFD in vertical farming has begun to intersect with disciplines such as urban engineering and bioclimatic architecture, thereby expanding the traditional scope of the field.

For its part, the most recent work by the group of [39], published in *Applied Thermal Engineering*, represents a stage of methodological maturity characterized by the integration of thermo-fluid phenomena, moisture transport, and experimental validation in real multi-layer systems. Although its citation count is lower due to its recent publication, the study introduces dimensionless uniformity indices and an integrated approach that links microclimate, energy efficiency, and crop growth quality. This type of contribution points toward the use of CFD not only as an analytical tool, but also as direct support for operational and energy-related decision-making in vertical farming systems. From a comparative perspective, the five articles share the use of three-dimensional CFD as their core methodological approach, but they differ in terms of scale of analysis, variables considered, and final objectives.

While earlier studies primarily focus on airflow velocity distribution, more recent works incorporate thermo-hygrometric variables, statistical uniformity metrics, and experimental validation. This progression suggests that scientific impact is closely linked to the ability of studies to integrate physical realism, practical applicability, and comparable metrics, thereby enhancing their reuse as references in subsequent research. In summary, the analysis of the top five most cited articles confirms that research on CFD applied to vertical and indoor agriculture has evolved from approaches centered on localized airflow improvements toward more systemic and integrated models, with productive, energetic, and urban implications. These works have not only established the methodological foundations of the field but have also outlined its main future trajectories, in which microclimate optimization, energy efficiency, and interaction with the built environment emerge as priority research axes.

Table 5: Most cited documents.

Document	Title	Citations	Reference	Journal
1	A CFD study on improving air flow uniformity in indoor plant factory system	120	[11]	Biosystems Engineering
2	A CFD analysis on improving lettuce canopy airflow distribution in a plant factory considering the crop resistance and LEDs heat dissipation	61	[15]	Biosystems Engineering
3	Analysis of climate uniformity in indoor plant factory system with computational fluid dynamics (CFD)	55	[16]	Biosystems Engineering
4	Effects of vertical farming on natural ventilation of residential buildings	34	[38]	Energy and Buildings
5	Thermo-fluid dynamic analysis of the air flow inside an indoor vertical farming system	23	[39]	Applied Thermal Engineering

Source: Authors, (2026).

III.13 KEYWORDS MOST FREQUENTLY USED BY DOCUMENT AUTHORS

From a quantitative and thematic perspective, the keyword cloud reveals a conceptual structure clearly dominated by the study of the air environment in controlled environment agriculture systems (Figure 8), once it is acknowledged that the high frequency of the term CFD (43 occurrences) directly reflects the focus of the search strategy. Excluding this methodological term, the most recurrent concepts are *Air* (19), *Plant Factory* (17), and *Ventilation* (15), indicating that a substantial proportion of the corpus concentrates on the characterization and control of airflow in plant factories and vertical farming systems [28], [40]. Likewise, the frequency of *Airflow* and *Crops* (8 occurrences each), together with *Cultivation* (7) and *Plant Growth* (6), reflects a balanced orientation between the optimization of physical variables and crop productive responses, suggesting that CFD is used not only to describe fluid dynamic phenomena but also to support decision-making related to agricultural performance [11], [28], [41].

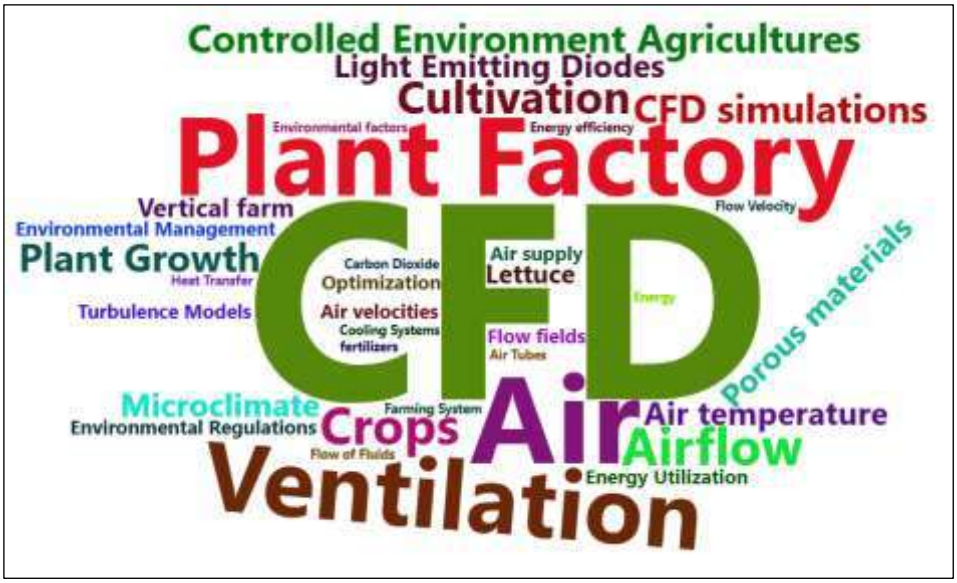


Figure 8: Word cloud of terms used by authors.
Source: Authors, (2026).

At a second level of analysis, the consistent appearance of medium-frequency terms such as *Microclimate*, *Air temperature*, *Light Emitting Diodes*, and *Porous materials* (5 occurrences each) indicates a growing interest in the interaction between ventilation, thermal conditions, artificial lighting, and structural components of the system, which are key elements for the integrated design of efficient indoor environments [42–44]. n parallel, the presence of keywords associated with *Energy Utilization*, *Optimization*, *Environmental Management*, and *Environmental Regulations* (3 occurrences) points to a gradual incorporation of energy efficiency, sustainability, and regulatory compliance criteria, although these aspects still hold a relatively limited weight within the field [45], [46]. Finally, the low frequency of terms such as *Carbon Dioxide*, *Heat Transfer*, and *Cooling Systems* (2 occurrences) suggests that these aspects, although recognized as relevant, remain underexplored, opening clear opportunities for future research aimed at more fully integrating thermal management, CO_2 concentration, and energy performance into CFD models applied to vertical and indoor agriculture [8], [47].

III.14 KEYWORD CO-OCCURRENCE NETWORK

The keyword co-occurrence network (Figure 9) is a bibliometric tool used to identify the conceptual structure and the main thematic axes of a research field based on the frequency with which specific terms appear jointly in scientific documents [48]. In this network, the nodes represent keywords, whose size reflects their relative relevance within the corpus, while the links indicate semantic relationships, that is, the co-occurrence of concepts within the same document [17]. Cluster analysis, based on the density and proximity of these links, makes it possible to group terms into coherent thematic subfields, facilitating the identification of dominant research lines, emerging areas, and interdisciplinary connections. In the context of applying Computational Fluid Dynamics (CFD) to controlled environment agriculture, this network is particularly useful for understanding how physical, biological, and energy-related variables are integrated in the study of microclimate and ventilation in indoor systems.

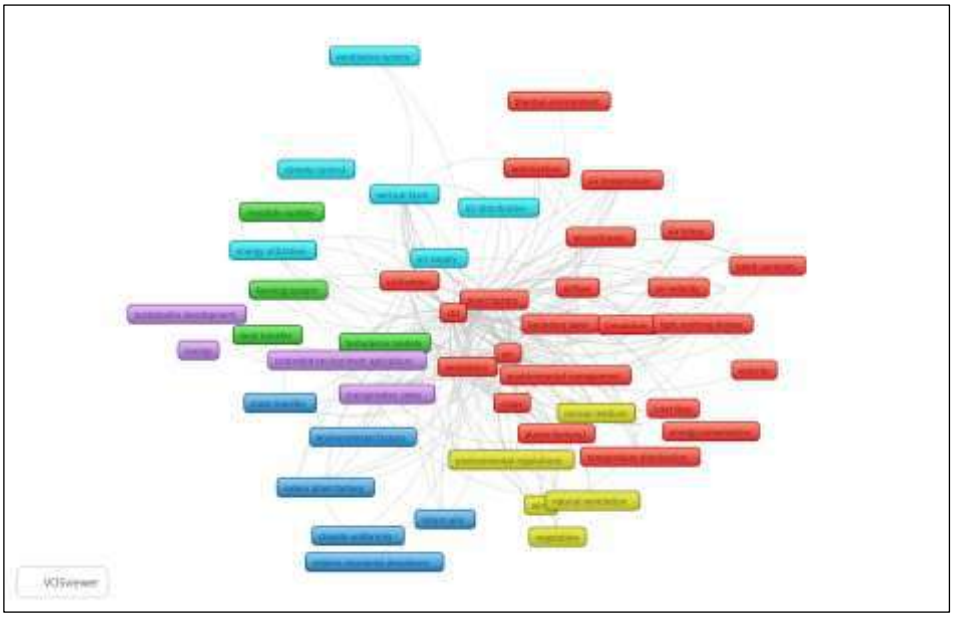


Figure 9: Keyword co-occurrence network of published documents.
Source: Authors, (2026).

Red Cluster: Microclimate, Ventilation, and Air Dynamics in Plant Factories: This cluster is the largest and densest in the network (30 items) and represents the core thematic nucleus of the field of study. It is composed of fundamental terms such as *air temperature*, *air velocity*, *airflow*, *boundary layer*, *ventilation*, *microclimate*, *plant factory*, *plant canopies*, *crops*, *cultivation*, *lettuce*, and *light emitting diodes*. The co-occurrence of these concepts indicates that research is strongly focused on the detailed analysis of air behavior and the thermal environment within plant factories, with particular emphasis on the interaction between airflow, the plant canopy, and artificial lighting [41], [43], [49–51]. The simultaneous presence of *CFD*, *simulation*, and *optimization* confirms that these phenomena are primarily addressed through advanced numerical modeling, aimed at optimizing microclimatic conditions to enhance crop growth and uniformity [29], [34], [52].

Green Cluster: Fluid Dynamics Fundamentals and Energy Transfer: This cluster groups terms such as *energy efficiency*, *heat transfer*, *turbulence models*, *Reynolds number*, *flow of fluids*, *experimental study*, and *indoor vertical farming*, representing the methodological and theoretical foundation of CFD applications. This cluster reflects research focused on the physical and numerical validation of models, as well as on the accurate representation of flow regimes and heat transfer mechanisms [31], [32], [53]. Although its size is smaller than that of the central cluster, its strong connection to the latter indicates that these studies provide essential scientific support to ensure the reliability of CFD simulations applied to indoor agricultural systems [28], [53].

Dark Blue Cluster: Climate Uniformity and Mass Transport: The cluster is dominated by terms such as *carbon dioxide*, *mass transfer*, *transport properties*, *climate uniformity*, *relative standard deviations*, *return airs*, and *indoor plant factory*. This thematic group focuses on the spatial distribution of gases and transport properties, with particular attention to microclimate homogeneity within closed cultivation systems [47], [54]. The presence of statistical indicators suggests an interest in quantifying internal environmental variability, a critical aspect for ensuring uniform crop growth and efficient resource use, but one that still appears as a less developed research line compared to ventilation and airflow studies [55], [56].

Yellow Cluster: Natural Ventilation and Semi-Porous: The cluster integrate concepts such as *natural ventilation*, *wind*, *porous medium*, *flow fields*, *vegetables*, and *environmental regulations*. This cluster represents a line of research oriented toward the interaction between external conditions and indoor systems, as well as the use of porous media models to represent the aerodynamic resistance of crops [38], [42]. Its presence indicates an effort to increase the realism of CFD models by incorporating passive ventilation conditions, regulatory constraints, and hybrid scenarios between closed and semi-open systems.

Purple Cluster: Sustainability, Energy, and Physiological Processes: The cluster groups terms such as *controlled environment*, *cooling systems*, *energy*, *fertilizers*, *sustainable development*, and *transpiration rates*. This cluster highlights an emerging dimension of the field related to sustainability, efficient energy use, and the integration of crop physiological processes into microclimate analysis [30], [57]. Although its relative weight is lower, its connection to the central cluster suggests a growing trend toward more holistic approaches, in which CFD is used not only to optimize airflow but also to assess energy and environmental impacts [45], [58].

Light Blue Cluster: Design and Operation of Ventilation Systems: The cluster includes terms such as *air distribution*, *air supply*, *ventilation system*, *energy utilization*, *climate control*, and *vertical farm*. This cluster acts as a bridge between CFD modeling and practical implementation, focusing on the design and operation of ventilation systems in vertical farming facilities [34], [40], [59], [60]. Its intermediate position within the network reflects its integrative role, linking fluid dynamic fundamentals with the operational management of indoor climate, which is essential for translating simulation results into real-world applications.

III.15 MULTIPLE CORRESPONDENCE ANALYSIS OF INTELLECTUAL STRUCTURE

Multiple Correspondence Analysis (MCA) is an exploratory statistical technique used to reduce the dimensionality of categorical data and to graphically represent the associations between variables and categories within a low-dimensional space. Unlike co-occurrence networks, MCA does not focus solely on the frequency of joint appearances, but rather on the geometric proximity between terms, allowing the identification of latent patterns, conceptual oppositions, and semantic groupings [61], [62]. In bibliometric studies, this technique is particularly useful for complementing network analyses, as it facilitates the interpretation of how concepts are organized around dominant conceptual axes [63].

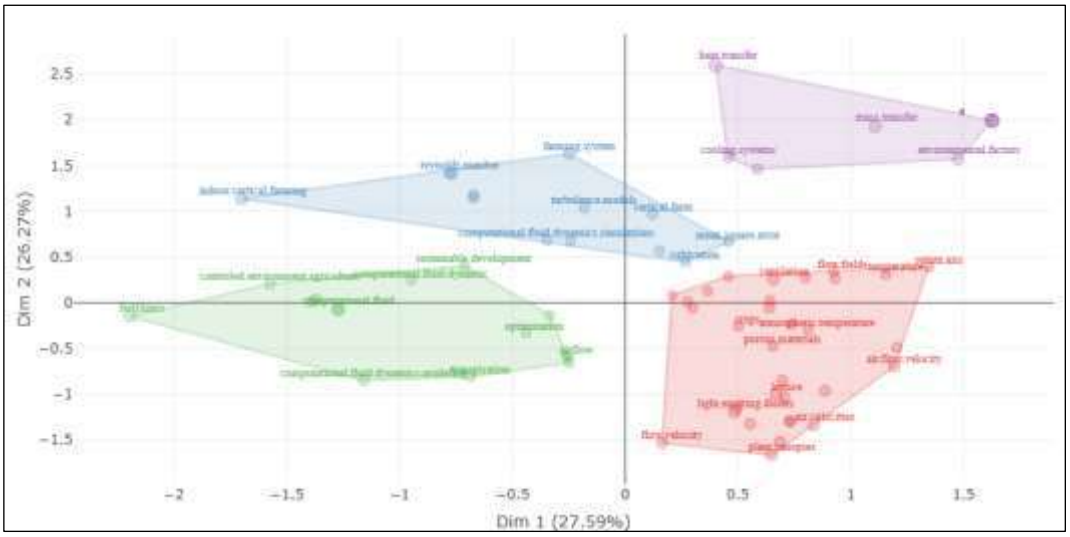


Figure 10: Analysis of dominant axes of intellectual structure analyzed.
Source: Authors, (2026).

In Figure 10 analyzed, the first two dimensions explain a substantial proportion of the total variability (Dim 1: 27.59% and Dim 2: 26.27%), indicating a well-defined conceptual structure. Dimension 1 clearly separates terms associated with the physical optimization of airflow and ventilation (right side), such as *airflow*, *air velocity*, *ventilation*, *porous materials*, and *light emitting diodes*, from those linked to more systemic and methodological approaches (left side), such as *computational fluid dynamics simulations*, *turbulence models*, *Reynolds number*, and *controlled environment agriculture*. This opposition suggests that the field is structured around studies focused on the design and operational performance of indoor systems versus works centered on the development and validation of numerical models.

Dimension 2, in turn, introduces a vertical differentiation between research focused on thermal and transfer management (*heat transfer*, *cooling systems*, *mass transfer*, *environmental factors*), located in the upper part of the plane, and studies more closely related to the plant canopy and air–plant interactions (*plant canopies*, *lettuce*, *air velocities*, *flow velocity*), situated in the lower part. This distribution indicates that, although both approaches are interconnected, there are differentiated research lines that prioritize, respectively, the energy–thermal control of the system and the microclimatic response at the crop scale. Overall, the MCA confirms the complementarity between engineering and biological approaches in the application of CFD to controlled environment agriculture, providing a structural perspective that reinforces and deepens the results obtained through keyword co-occurrence networks.

III.16 LIMITATIONS OF THE STUDY

This study presents several limitations inherent to both the bibliometric design and the information retrieval strategy. First, the search was conducted exclusively in the Scopus database, which, although it ensures high quality and multidisciplinary coverage, may underrepresent literature indexed in other databases (e.g., Web of Science, IEEE Xplore) as well as relevant grey literature for an applied field such as CFD in indoor agriculture. Second, the final corpus (52 documents) reflects a still emerging domain; therefore, some indicators (citations, local h-index, or network centrality) may be influenced by temporal window effects and by the uneven accumulation of citations between recent and foundational articles. In addition, the search equation itself includes methodological terms (CFD/ "computational fluid dynamics"), which may introduce a dominance bias of keywords such as CFD in co-occurrence analyses, reducing the relative visibility of substantive concepts such as CO_2, heat transfer, or VPD within the visualizations.

To address these limitations, several methodological decisions were adopted to strengthen the validity and traceability of the study. Regarding document coverage, a broad search strategy was implemented (with no restrictions by country or language), and a PRISMA-based workflow with explicit screening was applied (64 initial records; exclusion of 4 non-CFD studies and 8 due to thematic misalignment), enhancing transparency and reproducibility of the final corpus. With respect to citation and recency biases, results were interpreted by combining performance metrics (production, citations, indices) with structural analyses (bibliographic coupling, co-occurrence, and MCA), ensuring that the interpretation of the field did not rely solely on cumulative indicators. Finally, to reduce the dominance effect of the term "CFD" in semantic analyses, the interpretation of keyword clouds and networks explicitly differentiated between "methodological" terms (inherent to the search filter) and "thematic" terms, prioritizing the substantive analysis of clusters and conceptual relationships related to ventilation, microclimate, air distribution, energy, and mass transport.

III.17 FUTURE RESEARCH TRENDS AND CHALLENGES

One of the main emerging trends in research on the application of Computational Fluid Dynamics (CFD) in vertical and indoor agriculture is the transition from simplified models toward increasingly realistic and integrated representations of cultivation systems. Recent studies show a growing interest in incorporating detailed plant geometries, dynamic porous canopy models, and thermal sources associated with LED lighting, as well as in simultaneously integrating multiple microclimatic variables (temperature, humidity, air velocity, and CO_2 concentration). This evolution points to the use of CFD not merely as a descriptive tool, but as a predictive support for the comprehensive design and optimization of vertical farms, enabling the evaluation of complex operational scenarios prior to real-world implementation.

Nevertheless, significant methodological and conceptual challenges remain. Among these, the limited integration of crop physiological processes such as photosynthesis, transpiration, and stress responses into CFD models stands out, as these processes are still treated in a simplified or decoupled manner in many studies. In addition, critical variables such as the spatial distribution of CO_2, vapor pressure deficit (VPD), and latent heat transfer continue to be underrepresented compared to airflow analysis. This is compounded by the scarcity of full-scale experimental validations and longitudinal studies capable of assessing the microclimatic and energy performance of indoor systems over time, which restricts the generalizability of results and their industrial adoption

Looking ahead, research is expected to advance toward interdisciplinary and decision-oriented approaches, combining CFD with optimization techniques, artificial intelligence, and digital twins for the real-time control of vertical farms. The integration of energy analyses, sustainability assessments, and economic considerations represents another key axis for strengthening the practical applicability of CFD models. Furthermore, the reinforcement of international collaborations and the development of comparative studies across different indoor system typologies will help reduce regional gaps and consolidate a more robust knowledge framework. Overall, these trends and challenges outline a research agenda aimed at maximizing the productive, energy, and environmental efficiency of controlled environmental agriculture through the advanced use of CFD.

IV. CONCLUSIONS

From a quantitative perspective, the bibliometric analysis confirms that the application of Computational Fluid Dynamics (CFD) in vertical and indoor agriculture constitutes an emerging research field that is nevertheless in a clear phase of consolidation, with sustained growth in scientific production during the 2015–2025 period and a marked acceleration in recent years. The analyzed corpus, consisting of 52 documents published across 30 sources, reveals a high degree of geographical and editorial concentration, led mainly by China, the United States, and several European countries, as well as by journals specialized in biosystems engineering and controlled environment agriculture.

Bibliographic coupling and co-authorship networks display a highly centralized structure, supported by a relatively small core of authors and institutions with strong intellectual interconnections, indicating that the development of the field has been driven primarily by foundational contributions and recurrent collaborations rather than by a broad dispersion of approaches. From a conceptual standpoint, the results show that CFD has become a key tool for understanding and optimizing microclimate conditions in controlled environment agriculture systems, particularly with respect to ventilation, airflow distribution, and thermal uniformity.

The analysis of keywords, co-occurrence networks, and multiple correspondence analysis indicates that the thematic core of the field is structured around air–plant interactions, plant canopies, LED lighting, and energy efficiency, while variables such as CO_2 distribution, mass transport, and the integration of crop physiological processes remain at an early stage of development. This conceptual structure suggests a progressive evolution from approaches focused on localized airflow improvements toward more integrated models aimed at the simultaneous optimization of productive, energy, and environmental performance in vertical and indoor farming systems.

Looking ahead to future research, it is recommended to move toward more accessible, integrated, and decision-oriented CFD models, particularly in the context of low-cost indoor agriculture and in regions with limited resources. In this regard, priority should be given to developing approaches that combine simplified yet representative geometries, small-scale experimental validation, and clear metrics of uniformity and efficiency, enabling the transfer of generated knowledge to growers and designers of indoor systems with lower technological capacity.

In addition, integrating CFD with optimization tools, intelligent control strategies, and simplified energy analyses could facilitate the design of modular, scalable, and economically viable solutions. Finally, fostering international collaborations and comparative studies across different socioeconomic contexts will help broaden the impact of CFD as a strategic tool for improving the sustainability, resilience, and accessibility of controlled environment agriculture at a global scale

V. AUTHOR'S CONTRIBUTION

Conceptualization: Edwin Villagran, John Espita, Jorge Flores-Velasquez, Cruz Ernesto Aguilar, Jader Rodriguez.
Methodology: Edwin Villagran, John Espita, Jorge Flores-Velasquez, Cruz Ernesto Aguilar, Jader Rodriguez.
Investigation: Edwin Villagran, John Espita, Jorge Flores-Velasquez, Cruz Ernesto Aguilar, Jader Rodriguez.
Discussion of results: Edwin Villagran, John Espita, Jorge Flores-Velasquez, Cruz Ernesto Aguilar, Jader Rodriguez.
Writing – Original Draft: Edwin Villagran, John Espita, Jorge Flores-Velasquez, Cruz Ernesto Aguilar, Jader Rodriguez.
Writing – Review and Editing: Edwin Villagran, John Espita, Jorge Flores-Velasquez, Cruz Ernesto Aguilar, Jader Rodriguez.
Resources: Edwin Villagran, John Espita, Jorge Flores-Velasquez, Cruz Ernesto Aguilar, Jader Rodriguez.
Supervision: Edwin Villagran, John Espita, Jorge Flores-Velasquez, Cruz Ernesto Aguilar, Jader Rodriguez.
Approval of the final text: Edwin Villagran, John Espita, Jorge Flores-Velasquez, Cruz Ernesto Aguilar, Jader Rodriguez.

VI. REFERENCES

[1] E. Villagran, C. Bojacá, and M. Akrami, "Contribution to the Sustainability of Agricultural Production in Greenhouses Built on Slope Soils: A Numerical Study of the Microclimatic Behavior of a Typical Colombian Structure," Sustainability, vol. 13, no. 9, p. 4748, 2021.

[2] T. Kozai, C. Chun, and K. Ohyama, "Closed systems with lamps for commercial production of transplants using minimal resources," in XXVI International Horticultural Congress: Nursery Crops; Development, Evaluation, Production and Use 630, 2002, pp. 239–254.

[3] T. Kozai, "Plant factory in Japan-current situation and perspectives," Chron. Hortic, vol. 53, no. 2, pp. 8–11, 2013.

[4] L. I. Kun, F. Hui, and Z. O. U. Zhi-rong, "Optimization of rhizosphere cooling airflow for microclimate regulation and its effects on lettuce growth in plant factory," J. Integr. Agric., vol. 20, no. 10, pp. 2680–2695, 2021.

[5] E. Villagran et al., "Use of Lighting Technology in Controlled and Semi-Controlled Agriculture in Greenhouses and Protected Agriculture Systems—Part 1: Scientific and Bibliometric Analysis," Sustainability, vol. 17, no. 4, p. 1712, 2025.

[6] X. Gu and E. Goto, "Computational fluid dynamics model with realistic plant structures to study airflow in and around a plant canopy on a cultivation shelf in a plant factory with artificial light," Agriculture, vol. 14, no. 7, p. 1199, 2024.

[7] Y. Kitaya, "Importance of air movement for promoting gas and heat exchanges between plants and atmosphere under controlled environments," in Plant responses to air pollution and global change, Springer, 2005, pp. 185–193.

[8] E. Villagran et al., "CO2 Enrichment in Protected Agriculture: A Systematic Review of Greenhouses, Controlled Environment Systems, and Vertical Farms—Part 2," Sustainability, vol. 17, no. 7, p. 2809, 2025.

[9] H. Yu, H. Yu, B. Zhang, M. Chen, Y. Liu, and Y. Sui, "Quantitative perturbation analysis of plant factory LED heat dissipation on crop microclimate," Horticulturae, vol. 9, no. 6, p. 660, 2023.

[10] Z. H. Bian, Q. C. Yang, and W. K. Liu, "Effects of light quality on the accumulation of phytochemicals in vegetables produced in controlled environments: a review," J. Sci. Food Agric., vol. 95, no. 5, pp. 869–877, 2015.

[11] Y. Zhang, M. Kacira, and L. An, "A CFD study on improving air flow uniformity in indoor plant factory system," Biosyst. Eng., 2016, doi: 10.1016/j.biosystemseng.2016.04.012.

[12] B. Naranjani, Z. Najafianashrafi, C. Pascual, I. Agulto, and P.-Y. A. Chuang, "Computational analysis of the environment in an indoor vertical farming system," Int. J. Heat Mass Transf., vol. 186, p. 122460, 2022.

[13] Y. Haibo, Z. Lei, Y. Haiye, L. Yucheng, L. Chunhui, and S. Yuanyuan, "Sustainable development optimization of a plant factory for reducing tip burn disease," Sustainability, vol. 15, no. 6, p. 5607, 2023.

[14] G. Natarajan et al., “Modeling of air distribution inside a shipping container plant factory using computational fluid dynamics (CFD),” in AIP Conference Proceedings, 2022, vol. 2681, no. 1, p. 20091.

[15] H. Fang, K. Li, G. Wu, R. Cheng, Y. Zhang, and Q. Yang, “A CFD analysis on improving lettuce canopy airflow distribution in a plant factory considering the crop resistance and LEDs heat dissipation,” Biosyst. Eng., vol. 200, pp. 1–12, 2020.

[16] Y. Zhang and M. Kacira, “Analysis of climate uniformity in indoor plant factory system with computational fluid dynamics (CFD),” Biosyst. Eng., vol. 220, pp. 73–86, 2022.

[17] N. Montalván-Burbano, M. Pérez-Valls, and J. Plaza-Úbeda, “Analysis of scientific production on organizational innovation,” Cogent Bus. Manag., 2020, doi: 10.1080/23311975.2020.1745043.

[18] G. A. O. Rocha, M. A. Pichimata, and E. Villagran, “Research on the Microclimate of Protected Agriculture Structures Using Numerical Simulation Tools: A Technical and Bibliometric Analysis as a Contribution to the Sustainability of Under-Cover Cropping in Tropical and Subtropical Countries,” Sustain. 2021, Vol. 13, Page 10433, vol. 13, no. 18, p. 10433, Sep. 2021, doi: 10.3390/SU131810433.

[19] I. Passas, “Bibliometric Analysis: The Main Steps,” Encyclopedia, vol. 4, no. 2, 2024.

[20] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and P. Group*, “Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement,” Ann. Intern. Med., vol. 151, no. 4, pp. 264–269, 2009.

[21] C. Zhao, R. Geng, T. Chi, C. Khiewngamdee, and J. Liu, “Agricultural Technology Innovation and Food Security in China: An Empirical Study on Coupling Coordination and Its Influencing Factors,” Agronomy, vol. 14, no. 1, p. 123, 2024.

[22] L. Gurung, J. Rawal, G. Joshi, P. Rc, and A. Mandal, “Vertical Farming in Urban Agriculture: Opportunities, Challenges, and Future Directions,” Big Data Agric, vol. 6, no. 1, pp. 89–95, Jan. 2024, doi: 10.26480/bda.02.2024.89.95.

[23] A. Kluczkovski, P. Hadley, C. Yap, U. Ehgartner, B. Doherty, and K. Denby, “Urban vertical farming: innovation for food security and social impact?,” Philos. Trans. R. Soc. B Biol. Sci., vol. 380, no. 1935, 2025.

[24] J. J. Espitia et al., “CO2 Enrichment in Protected Agriculture: A Bibliometric Review on Greenhouses, Controlled Environment Systems, and Vertical Farms—Part 1,” Horticulturae, vol. 11, no. 5, p. 476, 2025.

[25] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, “How to conduct a bibliometric analysis: An overview and guidelines,” J. Bus. Res., vol. 133, pp. 285–296, 2021.

[26] B. H. Weinberg, “Bibliographic coupling: A review,” Inf. Storage Retr., vol. 10, no. 5–6, pp. 189–196, 1974.

[27] Y. Zhang and M. Kacira, “Analysis of environmental uniformity in a plant factory using computational fluid dynamics (CFD) analysis,” in International Symposium on New Technologies for Environment Control, Energy-Saving and Crop Production in Greenhouse and Plant 1227, 2017, pp. 607–614.

[28] Y. Zhang, C. Chen, H. Fang, and Y. Tong, “Improving Air Distribution Within Lettuce Plant Canopy by Employing Double-Channel Ventilation Cultivation System: Simulation and Experiment Study,” Agronomy, vol. 15, no. 10, p. 2326, 2025.

[29] L. Kang and T. van Hooff, “Numerical evaluation and optimization of air distribution system in a small vertical farm with lateral air supply,” Dev. Built Environ., vol. 17, p. 100304, 2024.

[30] W. Plas, T. Demeester, Z. Y. Ahmed, J. R. Coussement, K. Steppe, and M. De Paepe, “A detailed plant model in CFD that resolves the microclimate around individual leaves,” Biosyst. Eng., vol. 254, p. 104144, 2025.

[31] L. Kang, Y. Zhang, M. Kacira, and T. van Hooff, “CFD simulation of air distributions in a small multi-layer vertical farm: Impact of computational and physical parameters,” Biosyst. Eng., vol. 243, pp. 148–174, 2024.

[32] L. Kang and T. van Hooff, “On the improvement of air distribution performance in a vertical farm: Impact of airflow rate and baffles,” Therm. Sci. Eng. Prog., p. 103842, 2025.

[33] W. Plas and M. De Paepe, “Using CFD to improve flow conditions in vertical farms using realistic plant geometries,” in XXXI International Horticultural Congress (IHC2022): International Symposium on Advances in Vertical Farming 1369, 2022, pp. 49–56.

[34] L. Kang and T. van Hooff, “Numerical evaluation of an air distribution system in a generic multi-layer vertical farm: impact of air supply location,” in XXXI International Horticultural Congress (IHC2022): International Symposium on Advances in Vertical Farming 1369, 2022, pp. 57–64.

[35] G. A. O. Rocha, A. N. C. Medina, L. G. Arias, J. F. A. Caita, and E. Villagran, “Análisis sobre la actividad científica referente a las estrategias de climatización pasiva usada en invernaderos: Parte 2: análisis técnico,” Cienc. Lat. Rev. Científica Multidiscip., vol. 6, no. 6, pp. 2220–2245, 2022.

[36] D. A. Pinzon, G. M. Amado, J. Rodriguez, and E. Villagran, “Global Research Trends and Thematic Evolution of Blueberry (Vaccinium spp.) Science: A Bibliometric Analysis,” Horticulturae, vol. 11, no. 12, p. 1501, 2025.

[37] A. Bihari, S. Tripathi, and A. Deepak, “A review on h-index and its alternative indices,” J. Inf. Sci., vol. 49, no. 3, pp. 624–665, 2023.

[38] C. Yuan, R. Shan, A. S. Adelia, A. Tablada, S. K. Lau, and S. S.-Y. Lau, “Effects of vertical farming on natural ventilation of residential buildings,” Energy Build., vol. 185, pp. 316–325, 2019.

[39] G. Agati, B. Franchetti, F. Rispoli, and P. Venturini, “Thermo-fluid dynamic analysis of the air flow inside an indoor vertical farming system,” Appl. Therm. Eng., vol. 236, p. 121553, 2024.

[40] Z. Yu et al., “CFD-driven Bayesian optimization of localized ventilation to achieve desired microclimate conditions in plant factories,” Comput. Electron. Agric., vol. 240, p. 111225, 2026.

[41] A. M. Noh, M. A. M. Tahir, S. Mat, and M. H. Dzulkifli, “CFD simulation of temperature and airflow inside a shipping container size plant factory for optimal lettuce production,” Supplementary, vol. 4, no. S6, pp. 54–59, 2020.

[42] M. A. Sharief and A. Chowdhury, “Thermal computational analysis of microclimates for optimal crop production in controlled atmosphere,” Energy Procedia, vol. 160, pp. 783–790, 2019.

[43] X. Gu and E. Goto, “Evaluation of plant canopy microclimates with realistic plants in plant factories with artificial light using a computational fluid dynamics model,” Build. Environ., vol. 264, p. 111876, 2024.

[44] H. Fang, G. Wu, R. Cheng, and Y. Tong, “Numerical simulation of ventilation efficiency with perforated air tubes for controlling canopy environment in a plant factory based on CFD,” Trans. Chinese Soc. Agric. Eng., vol. 37, no. 22, 2021.

[45] P.-W. Hwang, C.-H. Chen, and Y.-J. Chang, “A study on energy strategy of a plant factory using sustainable energy combined with computational fluid dynamics simulation: An innovative practice of green information systems,” in 2017 Computing Conference, 2017, pp. 517–522.

[46] H. Fang, R. Cheng, Y. Tong, and K. Li, “Numerical simulation of the cooling efficiency of circular serrated water-cooled LEDs using CFD in plant factory,” Trans. Chinese Soc. Agric. Eng., vol. 37, no. 7, 2021.

[47] A. A. Ashnani et al., “Large Eddy Simulation of environmental impacts on mass transport in laboratory-scale vertical farm,” Int. J. Heat Mass Transf., vol. 256, p. 127964, 2026.

[48] E. Villagran, J. J. Espitia, F. A. Velázquez, A. Sarmiento, D. A. S. Velandia, and J. Rodriguez, “Artificial Intelligence in Solar-Assisted Greenhouse Systems: A Technical, Systematic and Bibliometric Review of Energy Integration and Efficiency Advances,” Technologies, vol. 13, no. 12, p. 574, 2025.

[49] H. Fang, Y. Zhang, G. Wu, K. Li, Y. Zhang, and Y. Tong, “Simulation of the effects of ventilation on canopy environment using cultivation bed with draft tube in a plant factory,” Trans. Chinese Soc. Agric. Eng., vol. 39, no. 16, 2023.

[50] Y. Zhang and M. Kacira, “Air distribution and its uniformity,” in Smart Plant Factory: The Next Generation Indoor Vertical Farms, Springer, 2018, pp. 153–166.

[51] P. Jiang, X. Yang, Y. Luo, L. Long, Y. Shi, and W. Hu, “CFD simulation of flow field and temperature field of different air flow circulation modes in vegetable aeroponics cultivation box,” Trans. Chinese Soc. Agric. Eng., vol. 35, no. 16, 2019.

[52] G. Larochelle Martin and D. Monfet, “High-density controlled environment agriculture (CEA-HD) air distribution optimization using computational fluid dynamics (CFD),” Eng. Appl. Comput. Fluid Mech., vol. 18, no. 1, p. 2297027, 2024.

[53] H.-T. Chen, Y.-C. Huang, S. Rashidi, B.-L. Chen, and W.-M. Yan, “Numerical and experimental studies on heat transfer characteristics and ventilation for indoor vertical farming system,” Therm. Sci. Eng. Prog., vol. 52, p. 102667, 2024.

[54] T. Demeester, W. Plas, and M. De Paepe, “A detailed CFD model of heat and mass transfer around plants,” 2023.

[55] D. Lee, E.-J. Choi, S.-M. Lee, and W. Oh, “Feasibility of a water-cooled LED system for climate uniformity in indoor plant factory using numerical analysis,” in X International Symposium on Light in Horticulture 1423, 2024, pp. 121–128.

[56] H. Gao, Z.-C. Tan, M. Yang, C.-P. Ma, Y.-F. Tang, and F.-Y. Zhao, “Microclimate Air Motion and Uniformity of Indoor Plant Factory System: Effects of Crop Planting Density and Air Change Rate,” Appl. Sci., vol. 15, no. 8, p. 4329, 2025.

[57] H. Sarailoo, L. Campbell, and H. Bougherara, “A Comprehensive Review on the Application of Computational Fluid Dynamics in Enhancing Indoor Vertical Farm Microclimate,” J. Biosyst. Eng., pp. 1–25, 2025.

[58] J. Flores-Velazquez, M. Akrami, and E. Villagrán, “The Role of Radiation in the Modelling of Crop Evapotranspiration from Open Field to Indoor Crops,” Agronomy, vol. 12, no. 11, p. 2593, 2022.

[59] J. Wang et al., “Design and simulation of the circulating air supply system for a ventilated wall-type plant factory based on the principle of uniform flow plate,” Trans. CSAE, vol. 39, pp. 213–221, 2023.

[60] Y. Wang et al., “Flow field characteristics of double circulation upward return air in the large-scale plant factory,” Trans. Chinese Soc. Agric. Eng., vol. 14, no. 16, 2025.

[61] A. A. M. Davidescu, M.-S. Florescu, L. C. Mosora, M. H. Mosora, and E. M. Manta, “A bibliometric analysis of research publications of the Bucharest University of Economic Studies in time of pandemics: Implications for teachers’ professional publishing activity,” Int. J. Environ. Res. Public Health, vol. 19, no. 14, p. 8779, 2022.

[62] D. M. Cretu and G. Grosseck, “A bibliometric analysis of romanian educational research in web of science: Trends, challenges, and opportunities for global integration,” Educ. Sci., vol. 15, no. 3, p. 358, 2025.

[63] J. J. Espitia et al., “Solar Energy Applications in Protected Agriculture: A Technical and Bibliometric Review of Greenhouse Systems and Solar Technologies,” Agronomy, vol. 14, no. 12, p. 2791, 2024.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1110-1119. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3270

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

AN APPROACH TO AUTOMATION IN SMART WASTE CLASSIFICATION BASED ON MACHINE LEARNING FOR EMBEDDED DEVICES

Pham Trung Kien[1]

[1]Vietnam Institute of Tropical Technology & Environmental Protection (VITTEP), Ho Chi Minh City, Vietnam.

[1]http://orcid.org/0009-0009-9872-9572

Email: phamtrungkien999@gmail.com

ARTICLE INFO

Article History
Received: January 9, 2025
Revised: January 20, 2026
Accepted: January 30, 2026
Published: February 28, 2026

Keywords:
Waste classification,
Embedded devices,
Smart trash bins,
Machine learning,
Raspberry Pi

ABSTRACT

This study aims to develop automated smart waste classification systems using machine learning deployed on resource-constrained hardware, such as Raspberry Pi, with the goal of integrating these models into the initial stage of waste sorting—that is, the trash bin. This study proposes two models: ConvNeXt and an improved CNN model, which is referred to as PrQ-CNN. The improved PrQ-CNN model utilizes pruning and quantization techniques to reduce its size. Both models were trained on a dataset comprising two classes: Non-recyclable and Recyclable. The data were split into a training set and a testing set with a 90:10 ratio. The models were integrated into a Raspberry Pi embedded computer to compare metrics such as model size, inference time, accuracy, missed detection rate (MDR), and false discovery rate (FDR). Experimental results show that the PrQ-CNN model achieved an accuracy of 92%, while its size was 12 times smaller than that of the original model before pruning and quantization. The ConvNeXt model achieved an accuracy of 98%. The results from the improved PrQ-CNN and ConvNeXt models were compared with VGG16, ResNet, EfficientNet, and MobileNet.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Waste pollution presents a significant challenge, particularly for developing tropical cities with high population density [1]. Smart cities around the world are currently focusing on solid waste management and treatment [2],[3]. Among these, solid waste sorting and recycling take the lead as major concerns [2]. According to data from the World Bank, global waste volume is projected to increase sharply, reaching 3.4 billion tons by 2050 [4], [5]. Globally, over 43% of solid waste is not properly managed through open burning, illegal dumping, and unregulated landfills [4], [5]. Additionally, the United Nations Environment Programme (UNEP) estimates that manual waste handling causes 15 – 20% of workplace injuries in the recycling sector [4], [5].

Landfilling also produces a large amount of methane, a potent greenhouse gas that contributes significantly to global warming [6]. Methane has a significantly higher heat-trapping potential than carbon dioxide, thereby exerting a stronger impact on global climate change [6]. In summary, if solid waste is not effectively managed, sorted, and recycled, it will lead to increased expenses, time consumption, and can cause waste pollution complications, resulting in consequences such as environmental pollution, impacts on public health, as well as worldwide problems like global warming and climate change. The objective of this study is to build compact machine learning models that can be deployed on resource-constrained hardware, or embedded computers like Raspberry Pi, and other devices, for automated solid waste classification into Recyclable and Non-recyclable types.

This study focuses on optimization, deployment, and evaluation of existing models for embedded waste classification applications. The study also aims to compare these models based on parameters such as performance and size. These hardware platforms can be easily integrated to build smart trash bins that can be deployed in public areas, helping to reduce production costs while introducing artificial intelligence technology into the very first step of waste classification when people discard their waste. Specifically, the waste treatment system model targeted by this study involves building trash bins with processors based on resource-constrained hardware platforms equipped with machine learning models.

Whenever people dispose of waste into these bins, the bins will use artificial intelligence technology to automatically classify the waste as Recyclable or Non-recyclable and separate it into the corresponding section for easier processing and recycling. Accordingly, the main research question addressed in this study is: how can machine learning-based waste classification models be optimized and evaluated to achieve an effective balance between accuracy, model size, and inference speed for deployment on resource-constrained embedded devices.

Contributions

(i) Developing machine learning models for solid waste classification, including a ConvNeXt model on the Kaggle Waste Classification dataset, achieving a high accuracy of 98%.
(ii) Proposing an improved Convolutional Neural Network (CNN) model utilizing pruning and quantization techniques to reduce model size by 12 times compared to the baseline, while maintaining accuracy, making it suitable for resource-constrained hardware and embedded devices.
(iii) Evaluating the models' performance based on accuracy, confusion matrix, and other performance metrics.
(iv) Providing an empirical, deployment-oriented comparison of multiple deep learning models on embedded hardware.

II. RELATED WORKS

According to [7], proposed developing a smart trash bin system capable of real-time waste level monitoring and efficient collection using YOLOv8 and YOLOv5 deep learning models to significantly reduce the number of parameters and processes, thereby lowering hardware requirements. Iman Ranjbar et al. proposed building a large and specialized dataset of plastic waste images, including four types of plastics commonly used in construction projects: ABS, HDPE, PS, and PVC [8]. They leveraged transfer learning with models trained and evaluated on advanced CNNs and Vision Transformers, including ResNet, ResNeXt, RegNet, and Swin Transformer, with Swin Transformer yielding the highest accuracy of 93.28% [8]. Mirna Castro Bello et al. evaluated and compared the effectiveness of four CNN models for waste classification using a custom dataset of 1883 Non-recyclable and Recyclable waste images on a Raspberry Pi 4 Model B with an HQ camera, achieving an average CPU utilization of 86% and an inference time of 1900 ms [9]. Proposed a hybrid method that combines instance segmentation deep learning models with morphological machine learning models to automate the classification and evaluation of decoration and renovation waste on a labeled dataset comprising 53,000 images [10].

The proposed method demonstrated remarkable accuracy, with a low relative error of 2.8% in the overall evaluation between models' predictions and manual sorting [10]. Evaluated the performance of AI models such as Mask R-CNN and YOLOv8, in enhancing plastic waste classification [11]. The models were assessed for accuracy, mean average precision (mAP), precision, recall, F1 score, and inference time, with hyperparameter tuning performed through grid search [11]. Results showed that Mask RCNN achieved an accuracy of 0.912 and mAP of 0.911, performing better than YOLOv8 in tasks requiring detailed segmentation, despite a longer inference time of 200–350 ms [11]. Conversely, YOLOv8, with an accuracy of 0.867 and mAP of 0.922, excelled in real-time scenarios due to its shorter inference time of 80–160 ms [11]. Presented the WasteIQNet model, a hierarchy-aware deep hybrid model, for detailed waste classification [12]. The proposed framework was trained and evaluated on the WEDR dataset, a curated collection of 360,000 images reflecting actual waste conditions across major landfill areas of Delhi [12]. Numerous other studies, such as those in [13-17], are also noteworthy.

These recent and in-depth studies demonstrate researchers' significant interest in the field of waste management and treatment, as well as practical use of new technologies like AI and ML in these areas. Based on the reviewed literature, it is evident that most recent studies primarily concentrate on either (i) multi-class waste classification using large-scale datasets or (ii) achieving high classification accuracy through deep learning models without imposing strict constraints on computational resources. In contrast, only a limited number of studies have systematically examined the trade-offs among model accuracy, model size, and inference speed when deploying waste classification models on resource-constrained embedded platforms, such as the Raspberry Pi. Furthermore, comprehensive comparative evaluations incorporating embedded-oriented performance metrics, including missed detection rate (MDR), false discovery rate (FDR), and on-device inference time, remain scarce in the existing literature.

III. MODELS AND METHODS

III.1 IDENTIFIED RESEARCH GAPS AND CHALLENGES

Based on the literature review, several research gaps can be identified. First, many existing studies prioritize accuracy while overlooking deployment feasibility on embedded hardware with limited memory and computational power. Second, comparisons across models are often conducted on different datasets, making direct evaluation difficult. Third, few studies report embedded-oriented performance indicators such as MDR, FDR, and real inference latency on low-power devices. These gaps motivate this study to focus on practical, deployment-oriented evaluations, with particular emphasis on the implementation of machine learning models for at-source waste classification on resource-constrained embedded systems.

III.2 MODELS AND METHODS

This study focuses on building a solid waste management and treatment system where waste sorting is performed from the very first step, as residents dispose of waste into bins. The trash bin will automatically classify the waste as Recyclable or Non-recyclable using a machine learning model. For such a system, one of the challenges is the limitations of hardware platforms for end-devices. In addition to high accuracy, the model must be sufficiently lightweight and optimized to be seamlessly integrated into these devices. Concretely, the objectives of this study include two main goals:

(i) Building machine learning models for Non-recyclable and Recyclable waste classification that can be applied in the design and manufacturing of smart trash bins for waste management and treatment systems.
(ii) Discussing and proposing the best model that can be applied on resource-constrained hardware platforms for integration into end-user devices, specifically smart trash bins

III.3 SMART TRASH BIN AND DATASET DESCRIPTION

The dataset used in this study is the Waste Classification data from Kaggle. This dataset comprises two components: Training and testing set, split at a 90:10 ratio. Each component consists of two classes: O (Other: Non-recyclable) and R (Recyclable). Figure 1(a) illustrates the structure of the smart waste bin, which employs an AI-based model to perform source-level waste classification. When waste is deposited into the bin, the AI system classifies it into Recyclable and Non-recyclable categories and subsequently controls the conveyor motor to direct the waste into the corresponding compartment. Figure 1(b) presents sample images of the two waste classes, namely Non-recyclable waste and Recyclable waste, contained in the dataset.



(a) Smart trash bin. (b) Images in the dataset.
Figure 1: Smart trash bin structure and sample images from the dataset.
Source: Authors, (2026).

To ensure fair comparison, all models were trained and evaluated using the same dataset split, input resolution (224×224), and evaluation metrics. In this study, data augmentation techniques were not applied during the training process. This design choice was made to ensure that the performance evaluation reflects the model capability under realistic deployment conditions, where waste images captured at the source often exhibit limited variability and are not artificially enhanced. The original class distribution of the dataset was preserved to simulate real-world waste generation scenarios, in which recyclable and non-recyclable waste categories are inherently imbalanced.

Maintaining this natural imbalance allows for a more faithful assessment of the model's robustness and practical effectiveness, particularly in minimizing misclassification risks under real deployment conditions. Additionally, cross-validation was not conducted in this study. Instead, a fixed train–validation–test split was adopted to reduce computational overhead and experimental complexity, which is consistent with the constraints of resource-limited embedded platforms. This approach also aligns with the study's primary objective of deployment-oriented evaluation, where repeatable inference performance, model size, and execution time on embedded hardware are prioritized over exhaustive statistical validation. The hardware specifications used in this study are presented in Table 1.

Table 1: Technical specifications of hardware.

No	Hardware	Technical specifications
1	Raspberry Pi 5	- CPU: ARM Cortex-A76 (ARM v8) 64-bit - Clock speed: 4 x 2.4 GHz - GPU: VideoCore VII (800 MHz) - RAM: 8 GB LPDDR4X (4267 MHz) - SD card: microSD
2	Kaggle GPU T4x2	- Architecture: Turing - CUDA Cores: 2,560 - Tensor Cores: 320 - GPU Memory: 16 GB GDDR6 - Memory Bandwidth: 320 GB/s

Source: Authors, (2026).

III.4 MODEL ARCHITECTURES

III.4.1 PrQ-CNN Model

CNN model: The Convolutional Neural Network model is a type of deep learning algorithm based on multi-layer neural networks that can identify, recognize, and classify objects, as well as detect and segment objects in given images [18].

CNNs typically consist of four types of layers: Convolutional, Pooling, Activation Function, and Fully Connected. The convolutional layer plays a crucial role in the overall CNN structure. It is a set of filters—or kernels—applied to the data. The width, height, and weights of each kernel are used to extract features from the input data. Weight values in the first kernel are randomly assigned but gradually become more informed by the training data. The pooling layer is used to reduce the dimensionality of feature maps while retaining the most important data [18]. A filter applies a pooling operation (max, min avg) by sliding across the input data in the pooling layer. Max pooling is most commonly used. A non-linear layer follows the convolution. Non-linearity allows the generated output to be altered or terminated. This layer is used to constrain or saturate the output [18].

Every type of activation function in every type of neural network serves the essential function of mapping inputs to outputs. The input value is calculated by summing the weighted inputs of the neuron and its bias (if any). This indicates that the activation function determines whether to activate a neuron in response to a given input by producing an appropriate output. Fully Connected (FC) layers, also known as dense layers, are used in neural networks, especially in deep learning. They are a type of neural network layer where every neuron in the layer is connected to every neuron in the preceding and succeeding layers. FC layers are typically deployed at the end of a neural network to perform classification or prediction tasks based on extracted specific features, helping the network learn complex associations between features [18].

Pruning: Pruning is the process of removing unnecessary or less important parts of a model to make it more optimized and faster without significantly reducing its accuracy or performance [19]. Pruning can be categorized into structured, unstructured, and semi-structured pruning. Structured pruning removes filters, channels, or layers to create structured sparsity patterns. In unstructured pruning, individual weight values without structural constraints are considered for removal. Semi-structured pruning acts as a hybrid approach combining aspects of both structured and unstructured pruning [19].

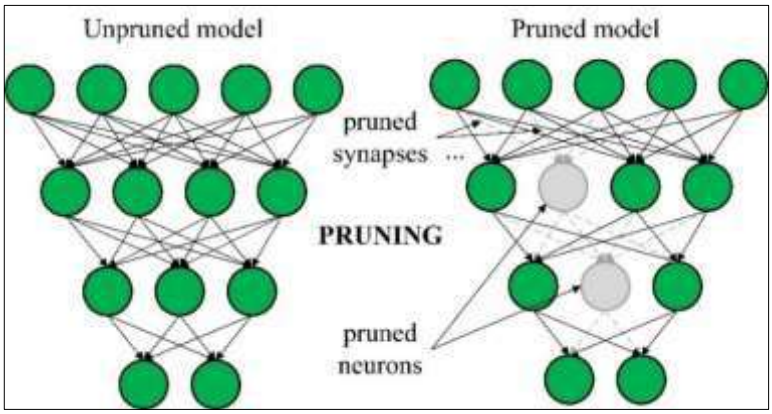


Figure 2: Pruning technique.
Source: Authors, (2026).

Quantization: Quantization is a compression method that converts high-precision representations (32-bit floating point) into lower-bit representations (int8 fixed-point) [20]. Quantization can reduce the CNN model's size, memory footprint, and power consumption, while also improving inference time. This is one of the deep learning model optimization techniques, particularly suitable for embedded systems, mobile devices, or applications requiring high performance and energy efficiency. Quantization techniques used for CNNs are broadly classified as: Post-Training Quantization and Quantization-Aware Training. CNN parameters need to be adjusted to maintain accurate performance after quantization. The process of model retraining to account for quantization is called Quantization-Aware Training, or it can be done without retraining through Post-Training Quantization [20].

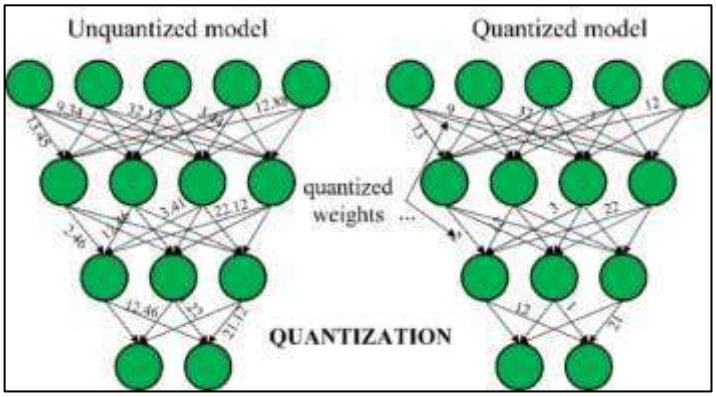


Figure 3: Quantization technique.
Source: Authors, (2026).

In this study, unstructured magnitude-based weight pruning was applied to the convolutional layers, in which individual weights with small absolute values were set to zero. This was followed by post-training dynamic range quantization, where the model weights were quantized to 8-bit integers while activations remained in floating-point precision. The pruning threshold was empirically selected to ensure that the accuracy degradation remained below 2%. Finally, the overall architecture of the PrQ-CNN is shown in Figure 4.

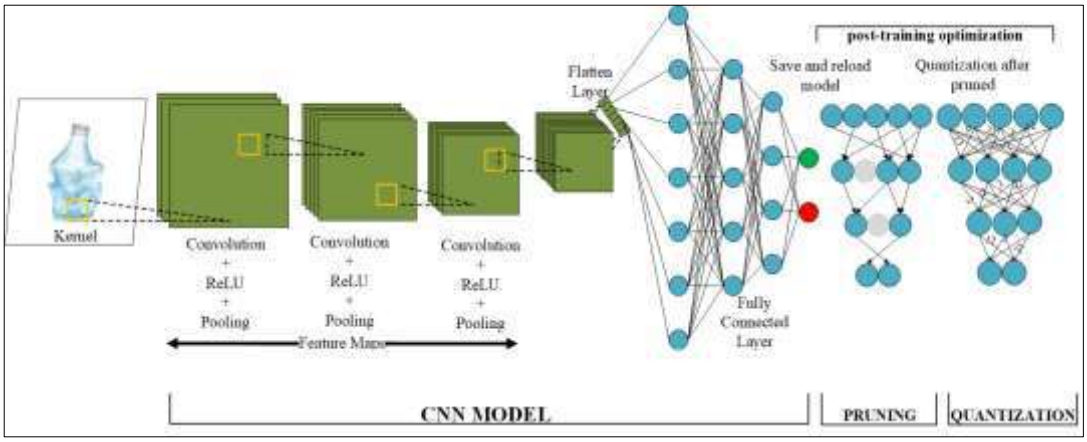


Figure 4: Proposed PrQ-CNN network architecture.
Source: Authors, (2026).

III.4.2 ConvNeXt Model

ConvNeXt is a deep neural network architecture designed to enhance the performance of image processing models, particularly in image recognition and classification [21].

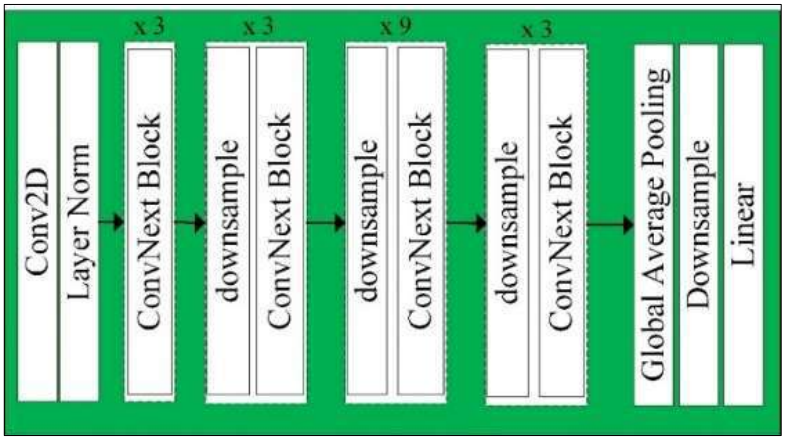


Figure 5: Architecture of the ConvNeXt (tiny) model used in this study.
Source: Authors, (2026).

It was introduced by the Facebook AI Research (FAIR) team in 2022 as a combination of features from traditional CNN models like ResNet and principles of Transformer architecture, aiming to optimize image processing capabilities [21].

IV. RESULTS AND DISCUSSION

IV.1 EVALUATION OF WASTE CLASSIFICATION MODEL RESULTS

To evaluate the performance of the models, this paper calculates and analyzes specific metrics as detailed below:

Confusion Matrix:

A confusion matrix is a specific tabular layout that allows for the visualization of an algorithm's performance. Each row of the matrix represents instances in an actual class, while each column represents instances in a predicted class, or vice versa.

Table 2: Confusion Matrix.

		Predicted Class	
		Yes	No
Actual Class	Yes	TP(a)	FN(b)
	No	FP(c)	TN(d)

Source: Authors, (2026).

Where:
(*) a: TP (true positive) – This occurs when both the actual class and the predicted class of the data point are True.
(*) b: FN (false negative) – This occurs when the actual class of the data point is True, but the predicted class is False (for example, when a pregnancy test indicates that a woman is not pregnant, but she actually is).
(*) c: FP (false positive) – This occurs when the actual class of the data point is False, but the predicted class is True (for example, when a pregnancy test indicates that a woman is pregnant while she is not).
(*) d: TN (true negative) – This occurs when both the actual class and the predicted class of the data point are False.

Accuracy:

Accuracy in classification is the most common evaluation metric. It is defined as the ratio of the number of correct predictions made by the model to the total number of predictions.

Accuracy

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} = \frac{a + d}{a + b + c + d} \quad (1)$$

Error Rate:

The Error Rate of the model is calculated as follows:

Error Rate

$$ErrorRate = 1 - Accuracy \quad (2)$$

Precision:

Precision is defined as the number of true positives divided by the sum of true positives and false positives. It indicates the proportion of positive predictions that are actual positives, in relation to the total number of predicted positives.

Precision

$$Precision = \frac{TP}{TP + FP} = \frac{a}{a + c} \quad (3)$$

Recall/Sensitivity:

Recall is defined as the number of true positives divided by the sum of true positives and false negatives. It indicates the proportion of positive predictions that are actual positives, relative to the total number of positives in the original dataset.

Recall/Sensitivity

$$Recall = \frac{TP}{TP + FN} = \frac{a}{a + b} \quad (4)$$

F1 Score:

The F1 score combines precision and recall for a specific positive class. It reflects the balance between precision and recall, particularly in cases of uneven class distribution. The F1 score achieves its highest value of 1 and its lowest value of 0.

F1 Score

$$F1Score = \frac{2 * Recall * Precision}{Recall + Precision} = \frac{2a}{2a + b + c} \quad (5)$$

MDR (Missed Detection Rate):

MDR is the percentage of fields that are actually positive but are missed (not detected).

MDR

$$MDR = \frac{FN}{TP + FN} = \frac{b}{a + b} \quad (6)$$

FDR (False Discovery Rate):

FDR is the proportion of false positives (incorrect positive detections) among all positive detections.

FDR

$$FDR = \frac{FP}{TP + FP} = \frac{c}{a + c} \quad (7)$$

These models were trained on the Kaggle platform using a T4x2 GPU. The results will be evaluated by implementing the testing set. The improved CNN model (PrQ-CNN) and ConvNeXt will be evaluated alongside VGG16, MobileNet, ResNet, and EfficientNet using performance parameters such as model size (MB), accuracy (%), average inference time per sample (ms), confusion matrix, and other parameters mentioned above on Kaggle GPU and Raspberry Pi. Specifically, the prediction results of the models are listed in Tables 3 to 8.

Table 3: Results of the PrQ-CNN prediction model.

	Precision	**Recall**	**F1-score**	**Support**
Non-recyclable	0.90	0.96	0.93	1401
Recyclable	0.94	0.87	0.91	1112
Accuracy			0.92	2513
Macro avg	0.92	0.91	0.92	2513
Weighted avg	0.92	0.92	0.92	2513

Source: Authors, (2026).

Table 4: Results of the ConvNeXt prediction model.

	Precision	Recall	F1-score	Support
Non-recyclable	0.97	0.99	0.98	1401
Recyclable	0.99	0.96	0.97	1112
Accuracy			0.98	2513
Macro avg	0.98	0.97	0.98	2513
Weighted avg	0.98	0.98	0.98	2513

Source: Authors, (2026).

Table 5: Results of the VGG16 prediction model.

	Precision	Recall	F1-score	Support
Non-recyclable	0.86	0.96	0.91	1401
Recyclable	0.94	0.80	0.87	1112
Accuracy			0.89	2513
Macro avg	0.90	0.88	0.89	2513
Weighted avg	0.90	0.89	0.89	2513

Source: Authors, (2026).

Table 6: Results of the MobileNet prediction model.

	Precision	Recall	F1-score	Support
Non-recyclable	0.85	0.97	0.90	1401
Recyclable	0.94	0.87	0.86	1112
Accuracy			0.88	2513
Macro avg	0.90	0.87	0.88	2513
Weighted avg	0.89	0.88	0.88	2513

Source: Authors, (2026).

Table 7: Results of the ResNet prediction model.

	Precision	Recall	F1-score	Support
Non-recyclable	0.88	0.98	0.92	1401
Recyclable	0.97	0.83	0.89	1112
Accuracy			0.91	2513
Macro avg	0.92	0.90	0.91	2513
Weighted avg	0.92	0.91	0.91	2513

Source: Authors, (2026).

Table 8: Results of the EfficientNet prediction model.

	Precision	Recall	F1-score	Support
Non-recyclable	0.88	0.98	0.92	1401
Recyclable	0.96	0.83	0.89	1112
Accuracy			0.91	2513
Macro avg	0.92	0.90	0.91	2513
Weighted avg	0.92	0.91	0.91	2513

Source: Authors, (2026).

The precision scores for both classes in the ConvNeXt model are the highest (0.97 and 0.99). In contrast, the other models exhibit significant discrepancies between the two classes, with the Non-recyclable class generally performing worse than the Recyclable class. Overall, the ConvNeXt model still demonstrated the best performance, as indicated by its F1-Score, reflecting good balance across both classes. This is followed by PrQ-CNN, which achieved an accuracy of 92%. The confusion matrices of the models are shown in Figure 6.

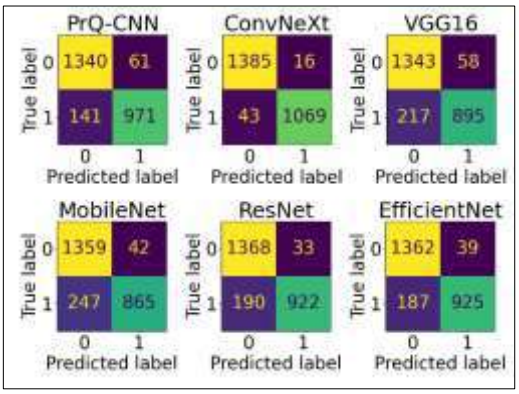


Figure 6: Confusion matrix of the models.
Source: Authors, (2026).

The ConvNeXt model exhibited the most exceptional performance with high accuracy, low error rate, and satisfactory positive class detection. PrQ-CNN followed with strong accuracy. ResNet and EfficientNet demonstrated comparable performance, both commendable but not superior to the ConvNeXt and PrQ-CNN models. The VGG16 and MobileNet models showed lower performance compared to the other models.

IV.2 DISCUSSION

The PrQ-CNN model is relatively small in size (22 MB) and fast, making it suitable for applications requiring quick processing and limited resources, such as embedded computers (92% accuracy). The ConvNeXt model achieved the highest accuracy (98%) but demands more resources than PrQ-CNN. It is ideal for applications where high accuracy is a priority. With an inference time of 234.3 ms/sample on Raspberry Pi, the ConvNeXt model still performs effectively when integrated into embedded devices. In general, the VGG16 and EfficientNet models are larger and slower compared to the other models. MobileNet is the lightest model (9.6 MB), making it suitable for real-time applications or devices with limited resources, though it sacrifices some accuracy, achieving 88% compared to other models. A comparison chart of the models based on six specific parameters is shown in Figure 8.

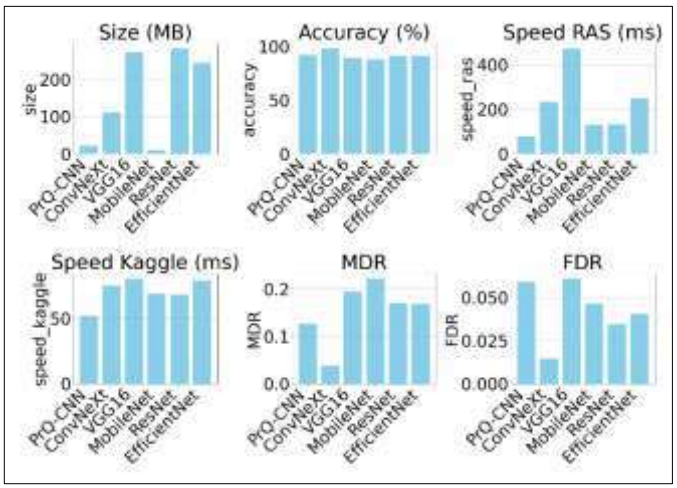


Figure 7: Comparison of models' parameters.
Source: Authors, (2026).

Tables 9 and 10 evaluate the performance of the models.

Table 9: Comparison table of best-performing models.

No	Parameter	Value	Model
1	Size	9.6 MB	MobileNet
2	Accuracy	98.0%	ConvNeXt
3	Speed on Raspberry Pi	76.3 ms	PrQ-CNN
4	Speed on Kaggle	52.0 ms	PrQ-CNN
5	MDR	0.0387	ConvNeXt
6	FDR	0.0147	ConvNeXt

Source: Authors, (2026).

Table 10: Comparison table of worst-performing models.

No	Parameter	Value	Model
1	Size	283.4 MB	ResNet
2	Accuracy	88.0%	MobileNet
3	Speed on Raspberry Pi	472.2 ms	VGG16
4	Speed on Kaggle	80.4 ms	VGG16
5	MDR	0.2221	MobileNet
6	FDR	0.0609	VGG16

Source: Authors, (2026).

Based on the two comparison tables above, several conclusions can be drawn: The ConvNeXt model achieves the best accuracy, while PrQ-CNN remains the leading model in terms of inference time on embedded computers. If accuracy is prioritized, ConvNeXt and PrQ-CNN are top choices. If a lightweight and fast-processing model for resource-constrained devices is required, PrQ-CNN is the most suitable. In general, PrQ-CNN provides a good balance between accuracy, size, and inference time for embedded devices. Table 11 compares the models proposed in this study with those reported in previous studies. The accuracy is evaluated on the same dataset as well as on different datasets. Overall, the ConvNeXt and PrQ-CNN models in this study demonstrated strong performance compared to the studies listed in Table 11. Both models exhibited superior accuracy, with PrQ-CNN leveraging pruning and quantization techniques to reduce the model size by approximately 12 times compared to the baseline unoptimized CNN model, making it suitable for embedded devices with limited hardware resources.

Table 11: Comparison with other models.

No	Study	Model	Dataset	Accuracy
1	This study	ConvNeXt	Kaggle	98.0%
2	This study	PrQ-CNN	Kaggle	92.0%
3	[22]	MLH-CNN	TrashNet	92.6%
4	[23]	EfficientNet	Custom Kaggle	85.0%
5	[24]	CNN	Same dataset	80.88%
6	[25]	CNN	Same dataset	91.32%
7	[26]	YOLOv7-tiny	Various sources	86.8%

Source: Authors, (2026).

V. FUTURE RESEARCH DIRECTIONS

This study primarily evaluates computational efficiency and accuracy. Power consumption analysis, robustness to varying lighting conditions, and real-world deployment in physical smart trash bins were beyond the current scope but remain essential future research directions. Therefore, in the future, the models will be tested on real-world datasets under conditions such as low light, image noise, and environmental factors. Subsequently, the classification system will be integrated into smart trash bins, and fully integrated embedded systems will be designed and tested, including sensors, processors, user interfaces, and data transmission systems for remote monitoring and control. Classification models will be coordinated with IoT systems to collect larger datasets, analyze user behavior, and optimize waste collection and sorting processes. Developing unsupervised or semi-supervised learning will reduce dependence on costly labeled data. Small-scale pilot systems will be deployed in real urban areas to evaluate practical performance, stability, and adaptability of the model. User feedback will be collected to refine the waste management and treatment system in smart cities.

VI. CONCLUSION

The machine learning models proposed in this study, namely ConvNeXt and PrQ-CNN, have demonstrated high accuracy in classifying Non-recyclable and Recyclable solid waste, achieving 98% and 92%, respectively. This study also proposed techniques such as pruning and quantization to reduce model size for integration onto embedded devices with limited hardware resources, such as Raspberry Pi. Experimental results show that both models have achieved not only superior classification efficiency but are also suitable for integration into smart trash bin systems. This study highlights the potential for application in waste management and treatment classification systems, contributing to improved efficiency, reduced errors, and cost optimization. Compared to conventional models such as VGG16, ResNet, or MobileNet, the proposed models in this study offer a good balance between accuracy, size, and inference time, meeting the requirements of embedded systems. It also demonstrates the potential for applying artificial intelligence technology in waste classification, which not only enhances processing efficiency but also promotes the development of smart, sustainable waste management systems, contributing to environmental protection initiatives against climate change and improving quality of life in urban areas.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Pham Trung Kien.
Methodology: Pham Trung Kien.
Investigation: Pham Trung Kien.
Discussion of results: Pham Trung Kien.
Writing – Original Draft: R. Pham Trung Kien.
Writing – Review and Editing: Pham Trung Kien.
Resources: Pham Trung Kien.
Supervision: Pham Trung Kien.
Approval of the final text: Pham Trung Kien.

VIII. DATA AVAILABILITY STATEMENT

Data supporting reported results can be downloaded at: *https://www.kaggle.com/techsash/waste-classification-data* (accessed on 09 January 2026).

IX. REFERENCES

[1] T. Kumari and A. S. Raghubanshi, "Waste management practices in the developing nations: challenges and opportunities," in Waste Management and Resource Recycling in the Developing World, Elsevier, 2023, ch. 33, pp. 773–797, doi: 10.1016/B978-0-323-90463-6.00017-8.

[2] A. Lakhouit, "Revolutionizing urban solid waste management with AI and IoT: A review of smart solutions for waste collection, sorting, and recycling," Results in Engineering, vol. 25, 2025, doi: 10.1016/j.rineng.2025.104018.

[3] A. Addas, M. N. Khan, and F. Naseer, "Waste management 2.0: Leveraging Internet of Things for an efficient and eco-friendly smart city solution," PLoS One, vol. 19, 2024, doi: 10.1371/journal.pone.0307608.

[4] F. Fotovvatikhah, I. Ahmedy, R. M. Noor, and M. U. Munir, "A systematic review of AI-based techniques for automated waste classification," Sensors, vol. 25, 2025, doi: 10.3390/s25103181.

[5] Z. Zhang, C. Zhonghao, J. Zhang, Y. Liu, L. Chen, M. Yang, A. I. Osman, M. Farghali, E. Liu, D. Hassan, I. Ihara, K. Lu, D. W. Rooney, and P. S. Yap, "Municipal solid waste management challenges in developing regions: A comprehensive review and future perspectives for Asia and Africa," Science of the Total Environment, vol. 930, 2024, doi: 10.1016/j.scitotenv.2024.172794.

[6] J. Ma, Z. Chen, J. Li, L. Liu, Q. Yang, Y. Wan, X. Lu, Y. Zhang, and X. Fei, “Concurrent aerobic methane oxidation and biodegradation of waste in shallow layer of landfill during aeration,” Journal of Environmental Management, vol. 381, 2025, doi: 10.1016/j.jenvman.2025.125282.

[7] M. M. Abo-Zahhad and M. Abo-Zahhad, “Real-time intelligent garbage monitoring and efficient collection using YOLOv8 and YOLOv5 deep learning models for environmental sustainability,” Scientific Reports, vol. 15, 2025, doi: 10.1038/s41598-025-99885-x.

[8] I. Ranjbar, Y. Ventikos, and M. Arashpour, “Deep learning-based construction and demolition plastic waste classification by resin type using RGB images,” Resources, Conservation and Recycling, vol. 212, 2025, doi: 10.1016/j.resconrec.2024.107937.

[9] M. C. Bello, D. B. E. Padilla, C. M. Morales, W. C. Francisco, C. V. M. Vega, C. M. Duarte, Y. E. Alcocer, and D. E. G. Valencia, “Convolutional neural network models in municipal solid waste classification: Towards sustainable management,” Sustainability, vol. 17, 2025, doi: 10.3390/su17083523.

[10] P. Wang, J. Xiao, R. Liu, X. Qiang, X. Duan, and C. Liang, “A hybrid classification and evaluation method based on deep learning for decoration and renovation waste in view of recycling,” Waste Management, vol. 191, pp. 1–12, 2025, doi: 10.1016/j.wasman.2024.10.027.

[11] J. Son and Y. Ahn, “AI-based plastic waste sorting method utilizing object detection models for enhanced classification,” Waste Management, vol. 193, pp. 273–282, 2025, doi: 10.1016/j.wasman.2024.12.014.

[12] S. Tiwari, S. Bisht, and K. Sharma, “Intelligent waste management using WasteIQNet with hierarchical learning and meta-optimization,” IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3574095.

[13] J. Tanvir, M. S. K. Tanzir, B. K. Paul, M. Morshed, “TrashNeXt: Classification of recyclable water pollutants using deep transfer learning method,” Case Studies in Chemical and Environmental Engineering, vol. 11, 2025, doi: 10.1016/j.cscee.2024.101073.

[14] G. Mazanov, A. Iliushina, S. Nesteruk, A. Pimenov, A. Stepanov, and N. Mikhaylova, “Hyperspectral data-driven solid waste classification,” IEEE Access, vol. 13, pp. 46659–46672, 2025, doi: 10.1109/ACCESS.2025.3551097.

[15] M. S. B. Islam, M. S. I. Sumon, M. E. Majid, S. B. A. Kashem, M. Nashbat, A. Ashraf, A. Khandakar, A. K. A. Kunju, M. Hasan-Zia, and M. E. H. Chowdhury, “ECCDN-Net: A deep learning-based technique for efficient organic and recyclable waste classification,” Waste Management, vol. 193, pp. 363–375, 2025, doi: 10.1016/j.wasman.2024.12.023.

[16] H. T. N. Le and H. Q. T. Ngo, “Application of the vision-based deep learning technique for waste classification using the robotic manipulation system,” International Journal of Cognitive Computing in Engineering, vol. 6, pp. 391–400, 2025, doi: 10.1016/j.ijcce.2025.02.005.

[17] A. M. Vukicevic, M. Petrovic, N. Jurisevic, M. Djapan, N. Knezevic, A. Novakovic, and K. Jovanovic, “Versatile waste sorting in small batch and flexible manufacturing industries using deep learning techniques,” Scientific Reports, vol. 15, 2025, doi: 10.1038/s41598-025-87226-x.

[18] M. M. Taye, “Theoretical understanding of convolutional neural networks: Concepts, architectures, applications, and future directions,” Computation, vol. 11, 2023, doi: 10.3390/computation11030052.

[19] J.-H. Luo, H. Zhang, H.-Y. Zhou, C.-W. Xie, J. Wu, and W. Lin, “ThiNet: Pruning CNN filters for a thinner net,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, 2019, doi: 10.1109/TPAMI.2018.2858232.

[20] S. I. Young, Z. Wang, D. Taubman, and B. Girod, “Transform quantization for CNN compression,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 9, pp. 5700–5714, 2022, doi: 10.1109/TPAMI.2021.3084839.

[21] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 11966–11976, doi: 10.1109/CVPR52688.2022.01167.

[22] C. Shi, C. Tan, T. Wang, and L. Wang, “A waste classification method based on a multilayer hybrid convolution neural network,” Applied Sciences, 2021, doi: 10.3390/app11188572.

[23] M. Malik, S. Sharma, M. Uddin, C. L. Chen, C. M. Wu, P. Soni, and S. Chaudhary, “Waste classification for sustainable development using image recognition with deep learning neural network models,” Sustainability, 2022, doi: 10.3390/su14127222.

[24] N. Nnamoko, J. Barrowclough, J. Procter, “Solid Waste Image Classification Using Deep Convolutional Neural Network,” Infrastructures, 2022, doi: 10.3390/infrastructures7040047

[25] Th. Vadivel, Kumar Barathi, G. Arulkumaran, M. B. Anand and C. Cherubini, “Wastewater Management Using a Neural Network-Assisted Novel Paradigm for Waste Prediction from Vermicomposting,” in Water, MDPI, 2024, vol. 33, issue. 23, doi: 10.3390/w16233450

[26] R. Alam, I. Darmawan, P. A. Alda, and F. M. N. Sugiartana, “WasteInNet: Deep learning model for real-time identification of various types of waste,” Cleaner Waste Systems, vol. 10, 2025, doi: 10.1016/j.clwas.2024.100198.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1120-1131. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3285

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

DYNAMIC COMBINED ECONOMIC EMISSION DISPATCH IN RENEWABLE ENERGY MICROGRIDS USING ZEBRA OPTIMIZATION ALGORITHM

Ram Prasad Kannemadugu[1], Y. Venkata Krishna Reddy[*2], Katabathina Venkata Subbaiah[3]

[1, 3]Department of EEE, Sri Venkateswara University, Tirupati, Andhra Pradesh, India.
[2]Department of EEE, Sri Venkateswara College of Engineering, Tirupati, Andhra Pradesh, India.

[1]https://orcid.org/0009-0009-8663-8884, [2]https://orcid.org/0000-0001-5913-3391, [3]https://orcid.org/0009-0001-8595-9557,

Email: ramprasad2102@gmail.com , *yvkrishnareddy36@gmail.com, saivenkybuddy@gmail.com

ARTICLE INFO

Article History
Received: January 11, 2025
Revised: January 20, 2026
Accepted: January 30, 2026
Published: February 28, 2026

Keywords:
Renewable Energy Sources,
Microgrids,
Combined Economic Emission Dispatch,
Dynamic Load,
Zebra Optimization Algorithm.

ABSTRACT

The increasing demand for energy necessitates the implementation of sustainable solutions that incorporate renewable energy sources (RES), specifically solar and wind. Nevertheless, the inherent variability of renewable energy sources (RES) presents a challenge in the optimization of power dispatch, particularly in islanded microgrids. The Combined Economic Emission Dispatch (CEED) problem in a three-unit operational system under dynamic load demand conditions is the subject of this study. Four scenarios are taken into account for the operation of the system such as Thermal-only; Thermal + Wind; Thermal + PV; and Thermal + Wind + PV. In order to guarantee system reliability and reduce both generation cost and emissions, a weighted sum approach is implemented in conjunction with the Zebra Optimization Algorithm (ZOA). This work had done using the MATLAB 2020 software. The proposed method effectively manages constraints such as ramp rate limits and thermal unit operating boundaries, demonstrating enhanced efficiency and adaptability in comparison to traditional techniques.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Power systems are tasked with the efficient management of energy production, transmission, and distribution to consumers.Given that energy demand has recently exceeded production capacity, existing power stations must be used to their maximum potential instead of constructing new facilities to accommodate the load. The economic load dispatch (ELD) model is the only model that accounts for load demand. This is the only model capable of ascertaining the distribution of power-generating units while conforming to constraints. The study [1] have used evolutionary algorithms to provide computer-aided solutions for the issues related to economic load dispatch. Despite the limitations given by the system, the objective is to generate the maximum feasible amount of electricity.

In [2] uses a combination of artificial intelligence and renewable energy distribution to optimize existing power networks for various applications. This enhancement may serve several purposes. In [3] they have developed a unique hybrid optimization technique that enhances energy efficiency and dispatch to address the issues related to the optimization of economic load dispatch. The paper [4] had examine the cost-reliability trade-off using sophisticated evolving intelligence algorithms developed for the integration of renewable energy sources. These algorithms were developed for research purposes.The multi-objective CEED problem leads to a decrease in both pollution emissions and fuel expenditures.

At Emission Dispatch, our primary objective is to decrease the volume of dangerous pollutants emitted by coal-fired power plants. The pollutants consist of carbon monoxide, sulfur dioxide, and nitrogen oxides. The study [5] developed an algorithm for real-coded chemical processes aimed at minimizing costs and emissions related to economic emission load dispatch. This was executed to attain the study's principal objective. In [6] they have conducted comprehensive ELD testing, revealing that search group optimization enhances convergence time and solution accuracy. These results underscore the significance of search group optimization. The study [7] had extend the idea to intelligent distribution networks, which oversee voltage stability and include virtual power plants fueled by renewableenergysources. Measures are implemented to fortify the power grid.

The adoption of ELD requires continuous dynamic demand integration to minimize costs, emissions, and grid instability. To ensure that power systems operate effectively and sustainably amidst varying load patterns, it is essential to use adaptable optimization strategies. The paper [8] enhanced grid stability with the use of real-time economic load dispatch and renewable energy curtailment. This was achieved despite the fairly complex nature of the data processing operation. Despite incurring substantial processing costs, Dragonfly successfully enhanced the accuracy of probabilistic ELD [9]. Despite considering parameter sensitivity, [10] successfully optimized the thermal-wind economic load dispatch using chaotic sine cosine. The study [11] enhanced ELD convergence by artificial ecosystem optimization. This was achieved despite requiring substantial processing effort. To address the coal problem, one should integrate wind power plants with coal-fired thermal power facilities. This technique maximizes sustainability, cost-effectiveness, and pollution reduction to their fullest potential.

The grid's enhanced resilience has led to an increased need for optimization frameworks proficient in managing unpredictability and dynamic load distribution. It is regrettable that [12] could not successfully augment their research, despite their efforts to enhance short-term hydrothermal dispatch to boost production. In paper [13] successfully used ELD with multi-fuel thermal and renewable energy sources, leveraging the Equilibrium Optimizer. This enabled him to save costs by adjusting the ELD parameters to get the intended outcome. This was implemented to achieve the goal of minimizing expenditures to the greatest extent feasible. The study [14] introduced new avenues of research in their study on renewable-integrated ELD optimization methodologies; nevertheless, they failed to test the scalability of their approaches prior to advancing their work. The use of photovoltaics in economic load dispatch seeks to reduce fossil fuel consumption, operational costs, and environmental impact.

Due to the unpredictability of the power source, optimization enhancements are necessary for photovoltaic (PV) output. These modifications are essential to guarantee the stability of the grid, the efficiency of the power distribution system, and the economic viability of the system amid variable load conditions. Although challenging to quantify, optimization strategies for CEED, whether using PV and wind or not, are successful, as shown in [15]. Despite the challenges in calculating them, some tactics prove effective. The study [16] discovered that solar sharing via PSO was both cost-effective and environmentally beneficial. Despite solar electricity being an intermittent energy source, this was the circumstance. In [17] endeavored to enhance the precision of their predictions by including photovoltaic and transmission losses into economic emission dispatch (EED). This was undertaken to enhance the precision of their forecasts. Solar and wind energy exemplify renewable energy sources that enhance the efficiency of electric vehicles (EVs), decrease fuel costs, promote sustainability, and mitigate pollution.

Due to the sporadic characteristics of renewable energy sources, the use of sophisticated optimization and energy storage technologies is essential for sustaining grid stability. Despite their complexity, hybrid systems are more economical and enhance energy security. Hybrid systems also reduce costs. In the study [18] devised a metaheuristic algorithm inspired by bats to enhance resource use efficiency. This idea addresses the fundamental difficulty associated with renewable energy. Due to the unpredictability of renewable resources, [19] assert that probabilistic modeling was used to account for this factor. The study [20] enhanced the efficiency of the task by analyzing CEED optimization approaches that included photovoltaic and wind generation systems. In [21] they have used an advanced pyrotechnics algorithm to optimize the combined dispatch of renewable energy sources. The intricacy of scheduling is further exacerbated by the use of solar and wind power prediction models, as shown in [22].

The study [23] are doing research on the integration of wind, solar, and hydropower to develop a more dependable hybrid dispatching system. Following an analysis of the system's cost and performance, [24] conclude their study on the financial implications of incorporating solar energy into the existing power grid. In [25] maximize CEED amidst fluctuating energy supply by using sophisticated hybrid algorithms, while simultaneously managing costs, emissions, and grid stability. This enables them to get effective optimization. To effectively integrate intermittent renewable energy sources such as solar, thermal, and wind power into Economic Load Dispatch (ELD), rigorous optimization is essential. The Zebra Optimization Algorithm (ZOA) effectively addresses multi-objective constraints, enabling it to enhance grid stability, decrease generation costs, and minimize emissions. Consequently, it can achieve these aims. Ultimately, it will provide optimal dispatch solutions, hence enhancing the endurance of hybrid power networks. This is now achievable due to its adaptive search capabilities.

II. PROBLEM FORMULATION

By superimposing the input-output properties of the boiler-turbine-generator system, the fuel cost function [25] is usually expressed as:

$$F_{C_{Total}} = \sum_{i=1}^{N} f_i(P_{gi}) = \sum_{i=1}^{N} (a_i + b_i P_{gi} + c_i P_{gi}^2) \quad \$/hr \tag{1}$$

By integrating the NOx, SOx, and COx emissions of a thermal power plant, the entire emission function [25] is depicted as:

$$E_T = \sum_{i=1}^{N} f_i(P_{gi}) = \sum_{i=1}^{N} (\alpha_i + \beta_i P_{gi} + \gamma_i P_{gi}^2) \quad kg/hr \tag{2}$$

Constructing a single objective function that concurrently achieve both goals is inequitable, since reducing emissions and decreasing gasoline costs may be inherently contradictory. This is due to the potential incompatibility of these two objectives. The Combined Economic Emission Dispatch (CEED) function may be articulated as follows [25]. The aim is achieved by integrating the Weighted Sum and Price Penalty Factor (PPF) techniques.

$$CEED_T = W * F_{C_{Total}} + (1 - W) * h * E_T \tag{3}$$

The PPF is employed to transform emissions into a cost equivalent value.

$$h_i = \frac{F_C(P_{gi,min})}{E_T(P_{gi,max})} \tag{4}$$

Incorporating RES like solar and wind power into the ELD problem requires that the system adhere to inequality constraints and stability constraints [25] to maintain power balance and equilibrium.

II.1 EQUALITY CONSTRAINTS

$$\sum_{i=1}^{NG} P_{gi} + \sum_{i=1}^{M} W_{gi} + \sum_{i=1}^{S} PV_{gi} = P_{Demand} + P_{loss} \tag{5}$$

II.2 INEQUALITY CONSTRAINTS

$$P_{gi,min} \leq P_{gi} \leq P_{gi,max} \tag{6}$$

$$0 \leq W_{gi} \leq W_{rgi} \tag{7}$$

$$0 \leq PV_{gi} \leq PV_{rgi} \tag{8}$$

III. PROPOSED OPTIMIZATION ALGORITHM

III.1 ZEBRA OPTIMIZATION ALGORITHM

The study [26] developed a metaheuristic inspired by the natural behaviours of zebras in herding, migration, and survival. This metaheuristic is inspired by the natural strategies used by zebras. The approach is referred known as the Zebra Optimization Algorithm (ZOA). The ZOA is intended to replicate the balance that zebras maintain between foraging for food in uncharted territories and evading hazardous environments. It is particularly adept at addressing multi-objective optimization challenges, shown by the Combined Economic Emission Dispatch (CEED) in hybrid power systems, a topic for which it was specifically developed.

Population Initialization: Generate an initial population of n zebras (candidate solutions):

$$X_i^0 = lb + rand(1, D) \cdot (ub - lb) \quad for\ i = 1,2,..,n \tag{9}$$

Fitness Evaluation: Evaluate each zebra's fitness using the CEED objective function

$$fit_i = CEED_T(X_i) \tag{10}$$

Global Best Update: Update global best P_Z if a better fitness is found

$$P_Z^t = argmin(fit_i^t) \tag{11}$$

Phase 1 – Foraging Behaviour (Exploration): Zebra moves toward the best-known position

$$X_i^{t+1} = X_i^t + r \cdot (P_Z^t - I \cdot X_i^t) \tag{12}$$

If new position yields better fitness, update:

$$If\ f(X_i^{t+1}) < f(X_i^t) \Rightarrow X_i^t = X_i^{t+1} \tag{13}$$

Phase 2 – Predator Avoidance Behaviour (Exploitation): Strategy 1 (Lion Attack – Escape Strategy), Zebra escapes with random disturbance decreasing over time:

$$X_i^{t+1} = X_i^t + R \cdot (2r - 1) \cdot (1 - \frac{t}{t_{max}}) \cdot X_i^t \tag{14}$$

Strategy 2 (Other Predators – Defensive Strategy): Move based on interaction with randomly selected zebra AZ

$$X_i^{t+1} = X_i^t + r \cdot (AZ - I \cdot X_i^t) \tag{15}$$

Boundary Handling: Ensure that the updated position remains within bounds

$$X_i^{t+1} = min(max(X_i^t + 1, lb), ub) \quad (16)$$

Convergence Check: Repeat steps until the stopping criterion is met

$$t = t + 1, stop\ if\ t \geq t_{max}\ or\ convergence\ achieved$$

III.2 STEPS FOR ZOA-BASED CEED OPTIMIZATION

1. **Problem Definition:** Define the CEED objective function incorporating generation cost, emission levels, and system constraints.
2. **Population Initialization:** Generate initial zebra positions representing candidate power dispatch solutions.
3. **Fitness Evaluation:** Evaluate each zebra's fitness based on economic cost, emission output, and constraint satisfaction.
4. **Exploration & Exploitation:** Use ZOA's dynamic movement strategy to explore new solutions or exploit known good regions, guided by the global best and peer interaction.
5. **Constraint Enforcement:** Apply generator output limits, ramp rate limits, and power balance conditions to maintain feasible solutions.
6. **Best Solution Update:** Track and update the best-performing solution after each iteration.
7. **Convergence Criterion:** Repeat the process until the termination condition is met (e.g., max iterations or acceptable error level).
8. **Result Analysis:** Analyze the optimized dispatch results to determine the most cost-effective and environmentally efficient strategy.

IV. RESULTS AND DISCUSSIONS

This paper addresses the Combined Economic and Emission Dispatch (CEED) problem for a three-unit thermal power system under diverse integration scenarios of renewable energy sources. The system parameters, renewable generation profiles, and load demand are derived from [25]. The CEED objective function is constructed as a weighted sum of fuel costs and emission levels. The Zebra Optimization Algorithm (ZOA) is executed in MATLAB 2020 to address the dispatch problem in four scenarios: (i) thermal-only, (ii) thermal + PV, (iii) thermal + wind, and (iv) thermal + PV + wind. The outcomes are juxtaposed with benchmark solutions found in the current literature to authenticate the performance and effectiveness of ZOA. The ZOA model parameters are: R=0.1 (escape step size), r1 and r2 ∈ [0,1], and a dynamic equilibrium between exploration and exploitation based on iteration progress.

IV.1 CASE STUDY 1: EXCLUSIVE THERMAL UNITS

The CEED issue is addressed using the ZOA, focusing just on three thermal power units without including renewable energy sources. The aim is to reduce fuel expenses and pollution levels within a 24-hour timeframe, while maintaining power equilibrium and complying with operating restrictions. The system load requirement fluctuates between 140 MW and 250 MW during the day. According to the ideal dispatch plan established by ZOA, Unit 1 (Pg1) regularly functions at 37 MW, whereas Unit 3 (Pg3) sustains a stable output of 50 MW, suggesting that technical or economic considerations favour stationary operation. Unit 2 (Pg2) modulates its output between 53 MW and 160 MW, contingent upon the system's real-time load demands.

This guarantees a balanced alignment of supply and demand over the scheduling period. The optimization findings indicate a total fuel cost of 170,476 $/hr and an emission level of 3,711.05 kg per hour. The total CEED cost for the 24-hour period is 151,475 $/hr. The hourly CEED cost varies with demand, from a minimum of 5,395.10 $/hr during low-demand times to a high of 7,116.70 $/hr during peak load. These findings validate the efficacy of ZOA in addressing multi-objective optimization for thermal-only systems and provide a benchmark for evaluating enhancements from renewable integration in future scenarios.

Table 1: Hourly Optimal Results by ZOA for Case 1 (Thermal Only).

Hour	P1 (MW)	P2 (MW)	P3 (MW)	P-Thermal (MW)	Fuel-Cost ($/hr)	Emission (Kg/hr)	CEED ($/hr)
1	37	53	50	140	6051.1	112.85	**5395.1**
2	37	63	50	150	6256.2	116.13	5571.1
3	37	67.999	50	155	6359	118.37	5661.4
4	37	72.999	50	160	6461.9	121.01	5753.3
5	37	77.999	50	165	6565	124.05	5846.8
6	37	82.997	50	170	6668.2	127.49	5941.9
7	37	88	50	175	6771.6	131.33	6038.5
8	37	92.998	50	180	6875.1	135.57	6136.7
9	37	123	50	210	7499.3	169.41	6758.8
10	37	143	50	230	7918.5	199.97	6988.5
11	37	153	50	240	8128.9	217.64	**7116.7**
12	37	160	50	250	8340.1	229.45	7258.1
13	37	153	50	240	8128.9	217.65	**7116.7**
14	37	133	50	220	7708.6	183.89	6868.7
15	37	113	50	200	7290.6	156.53	6545.1

16	37	92.999	50	180	6875.1	135.57	6136.7
17	37	83	50	170	6668.2	127.49	5941.9
18	37	98	50	185	6978.7	140.21	6236.4
19	37	113	50	200	7290.6	156.53	6545.1
20	37	153	50	240	8128.9	217.64	**7116.7**
21	37	138	50	225	7813.5	191.72	6927.4
22	37	103	50	190	7082.5	145.25	6337.7
23	37	72.997	50	160	6461.9	121.01	5753.4
24	37	57.999	50	145	6153.6	114.29	5482.3
Total					**170476**	**3711.05**	**151475**

Source: Authors, (2026).

In Case 1, the whole demand for 24 hours is fully satisfied by thermal units, without any input from renewable sources. The thermal power production precisely corresponds to the hourly demand. In the first hour, the demand is 140 MW, entirely met by PT at 140 MW. At hour 10, demand reaches its zenith at 230 MW, with thermal generation precisely aligning at 230 MW. During peak demand hours, namely hour 12 (250 MW) and hour 20 (240 MW), the whole load is managed only by thermal plants, devoid of any renewable support. This configuration exemplifies the most conventional operation, yielding the greatest fuel expenses and emissions across all scenarios owing to exclusive dependence on fossil fuel-based generating. The whole power balance situation is shown in Table 6.



Figure 1: Hourly Variation of Fuel Cost and CEED – Case 1.
Source: Authors, (2026).

Figure 1 depicts the hourly fluctuations of Fuel Cost and CEED (Cost of Energy and Emission Dispatch) across a 24-hour duration. The horizontal axis indicates time in hours (1–24), and the vertical axis signifies fuel expense in $/hr. The blue curve represents Fuel Cost, whereas the orange curve illustrates CEED. Both measures have a same trajectory, with values rising throughout the morning hours and culminating at Hour 11, when Fuel Cost attains a high of $8340.10 and CEED nears $7298.10. Subsequent to this apex, both curves exhibit a slow descent, followed by a secondary ascent at Hour 21, signifying another operational high. The proximity of the two curves indicates a robust link between fuel use and emission-related expenses. Every data point on the graph is labelled for clarity, offering comprehensive insight into the fluctuations in operational costs throughout the day.

IV.2 CASE STUDY 2: THERMAL + PHOTOVOLTAIC(PV) SYSTEM

In Case 2, a 40 MW photovoltaic (PV) system is included into the power grid in conjunction with the three existing thermal units. The PV source, deemed non-dispatchable and devoid of emissions, functions with a stable generation profile that directly mitigates a portion of the overall system demand. This integration significantly decreases the net load on thermal generators, so alleviating their operating strain and enhancing overall system efficiency. The ZOA is used to address the modified CEED issue in accordance with the altered load profile. Following the integration of PV power, the thermal units Pg1 and Pg3 maintain their base load outputs of 37 MW and 50 MW, respectively, while Pg2 modulates its output to align with the residual demand.

The optimum dispatch plan embodies this tendency, allowing the system to sustain viability amid fluctuating demand levels. Significant reductions are achieved in both economic and environmental parameters relative to Case 1. The aggregate fuel expenditure diminishes to $166,659.86 per hour, emissions decline to 3,656.245 kg per hour, and the optimal CEED value is lowered to $148,343.14 per hour. The findings underscore the substantial benefits of integrating solar energy into traditional thermal systems, illustrating ZOA's efficacy in reducing costs and emissions in hybrid power systems.The incorporation of a 40 MW photovoltaic system in Case 2 alleviates the need on thermal production, particularly during daylight hours.

At hour 8, with demand at 180 MW, photovoltaic (PV) systems provide 16.18 MW, reducing the thermal load to 163.82 MW. At hour 10, photovoltaic (PV) systems reach their maximum production of 39.37 MW against a demand of 230 MW, hence decreasing thermal generation to 190.63 MW. In hour 13, photovoltaic (PV) production provides 31.94 MW, whereas thermal output has reduced to 208.06 MW, as seen in Table 6. During the early or late hours when sunshine is absent (e.g., hour 1 and hour 24), photovoltaic production ceases, and heat generation only fulfils the requirement. This partial integration mitigates fuel use and emissions during peak solar hours, hence improving overall efficiency.

Table 2: Hourly Optimal Results by ZOA for Case 2 (Thermal+PV).

Hour	P1 (MW)	P2 (MW)	P3 (MW)	P-Thermal (MW)	Fuel-Cost ($/hr)	Emission (Kg/hr)	CEED ($/hr)
1	37	53	50	140	6051.145	119.846	5395.138
2	37.0001	62.9991	50.0008	150	6256.2266	123.1261	5571.1
3	37	67.9995	50.0005	155	6358.991	125.366	5661.432
4	37	72.9994	50.0006	160	6461.9061	128.006	5753.338
5	37.0001	77.9998	50	165	6564.9701	131.0458	5846.809
6	37.0002	82.9698	50	169.97	6667.5654	134.4639	5941.279
7	37.0001	81.729	50.001	168.73	6641.956	133.5742	5917.569
8	37.0001	76.8197	50.0002	163.82	6540.6336	130.2924	5824.609
9	37.0002	98.9452	50.0046	185.95	6998.4514	148.131	6255.565
10	37.0004	103.625	50.0043	190.63	7095.6439	152.907	6350.636
11	37.0001	145.589	50.0007	232.59	7972.9227	211.3907	7020.94
12	37.0001	159.349	50.0005	246.35	8262.9119	236.7025	7201.801
13	37.0001	121.059	50.0004	208.06	7458.7749	173.7816	6716.816
14	37	106.186	50.0035	193.19	7148.8634	155.6689	6403.223
15	37.0004	102.918	50.0013	189.92	7080.8852	152.161	6336.111
16	37.0002	87.6995	50.0003	174.7	6765.3445	138.084	6032.63
17	37.0001	73.4279	50.0021	160.43	6470.7667	128.2518	5761.323
18	37.0002	95.6898	50	182.69	6930.8182	145.0124	6190.13
19	37.0002	112.999	50.0002	200	7290.6254	163.5255	6545.106
20	37.0002	152.999	50	240	8128.945	224.6456	7116.678
21	37	137.999	50.0008	225	7813.4513	198.7251	6927.431
22	37	102.999	50.0002	190	7082.5453	152.2459	6337.742
23	37.0001	72.9996	50.0004	160	6461.9057	128.006	5753.336
24	37	57.9998	50.0002	145	6153.6104	121.2861	5482.333
Total					**166659.859**	**3656.245**	**148343.1**

Source: Authors, (2026).

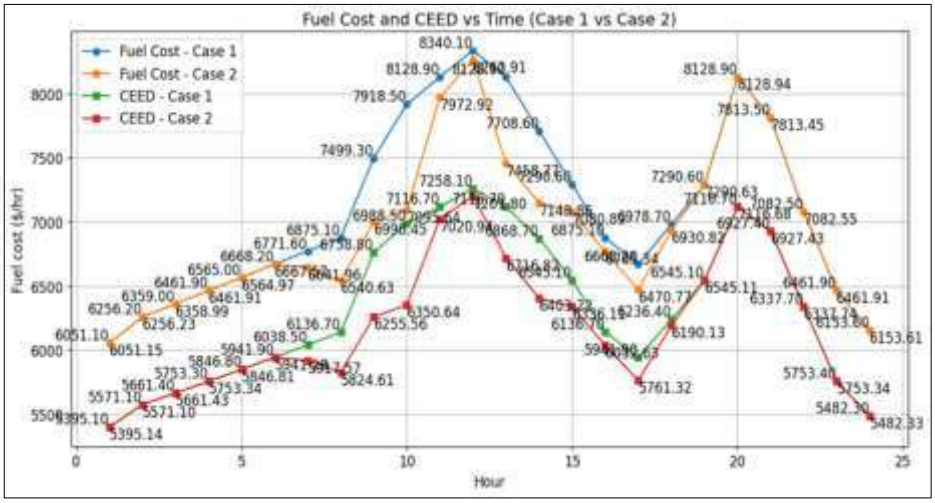


Figure 2: Hourly Variation of Fuel Cost and CEED – Case 1 and Case 2.
Source: Authors, (2026).

Figure 2 depicts the hourly fluctuations of Fuel Cost and CEED (Combined Economic and Emission Dispatch) for Case 1 and Case 2 across a 24-hour duration. The x-axis denotes the hours of the day, whilst the y-axis signifies the associated expenditures in $/hr. The graph delineates four curves: Fuel Cost-Case 1,Fuel Cost–Case 2,CEED–Case 1 and CEED-Case 2. Overall, Case 2 demonstrates elevated fuel expenses compared to Case 1 at peak demand intervals, particularly between hours 10–13 and 20–21, indicating a heightened operational strain. Nonetheless, the CEED values in Case 2 are decreased at certain hours, indicating a successful decrease in emissions or enhanced incorporation of cleaner energy sources such as solar (PV) systems. The fuel cost and CEED trends reach their zenith at Hour 11, signifying a pivotal period for operational planning. Case 2 illustrates an enhanced balance between economic dispatch and environmental consequences.

IV.3 CASE STUDY 3: THERMAL + WIND SYSTEM

In Case 3, a 30 MW wind power plant is incorporated into the existing thermal generating infrastructure. Similar to PV energy, wind power is a clean, non-dispatchable source; but, its intrinsically stochastic characteristics inject more unpredictability into the dispatch strategy. The CEED issue has been revised to include the dynamic wind profile, and the ZOA is used to ascertain the ideal generating schedule. The dispatch findings, as shown in Table 3, indicate that thermal units adjust to the fluctuations of wind power. Unit Pg1 operates consistently at its baseline of 37 MW, enhancing system stability. Pg3 functions consistently, while Pg2 adopts a flexible approach, modulating its output according to variations in wind availability. This architecture allows the system to effectively balance demand despite the unpredictability of renewable inputs.

In contrast to the other situations, Case 3 attains a total fuel cost of $166,041.9 per hour and emissions of 3,612.80 kg per hour, culminating in a CEED value of $147,886.74 per hour. The cost decrease in the PV-integrated scenario (instance 2) is minimal; nonetheless, this instance demonstrates increased fluctuation in emissions owing to the intermittent characteristics of wind energy. Nevertheless, the findings affirm that the incorporation of wind power improves both economic and environmental conditions compared to the thermal-only baseline. This scenario further illustrates ZOA's flexibility in controlling dispatch under uncertainty, showcasing its resilience in hybrid systems with dynamic renewable production.

Table 3: Hourly Optimal Results by ZEBRA Algorithm for Case 3 (Thermal+Wind).

Hour	P1 (MW)	P2 (MW)	P3 (MW)	P-Thermal (MW)	Fuel-Cost ($/hr)	Emission (Kg/hr)	CEED ($/hr)
1	37	51.2999	50	138.3	6016.341	117.4475	5365.85
2	37	54.4999	50.0001	141.5	6081.869	118.236	5421.132
3	37	58.7292	50.0008	145.73	6168.584	119.5299	5495.198
4	37	56.34	50	143.34	6119.575	118.7635	5453.209
5	37.0001	70.779	50.0009	157.78	6416.194	124.7845	5712.34
6	37.0001	78.0885	50.0014	165.09	6566.829	129.1043	5848.514
7	37	73.34	50	160.34	6468.909	126.2	5759.641
8	37.0001	67.4395	50.0005	154.44	6347.474	123.0952	5651.237
9	37	102.419	50.0003	189.42	7070.495	149.6407	6325.908
10	37	125.146	50.0038	212.15	7544.256	177.3812	6781.639
11	37.0001	140.199	50.0005	227.2	7859.639	200.3016	6954.016
12	37.0002	144.348	50.001	231.35	7946.846	207.2579	7005.336
13	37.0002	138.649	50.0003	225.65	7827.094	197.7738	6935.233
14	37	122.649	50.001	209.65	7491.993	173.9274	6751.173
15	37.0002	104.739	50.0002	191.74	7118.708	152.0933	6373.371
16	37.0001	79.2882	50.0018	166.29	6591.588	129.8951	5871.19
17	37.0001	79.5593	50.0007	166.56	6597.158	130.0763	5876.298
18	37	96.1297	50.0003	183.13	6939.942	143.4237	6198.922
19	37.0001	112.249	50.0007	199.25	7274.999	160.6239	6529.338
20	37.0001	152.829	50.001	239.83	8125.363	222.3307	7114.45
21	37.0003	137.847	50.0021	224.85	7810.305	196.4822	6925.656
22	37.0002	102.689	50.0007	189.69	7076.105	149.9213	6331.415
23	37.0001	71.9299	50	158.93	6439.869	125.4073	5733.535
24	37.0001	57.4195	50.0005	144.42	6141.717	119.0986	5472.14
Total					**166041.9**	**3612.796**	**147886.7**

Source: Aurthors, (2026).

Case 3 presents a 30 MW wind power system, yielding fluctuating but significant contributions that mitigate heat production. During the third hour, wind generation yields 9.27 MW in response to a demand of 155 MW, resulting in a reduction of PT to 145.73 MW. During windy periods, namely during hour 8, wind generation reaches a high of 25.56 MW, while power transmission decreases to 154.44 MW to meet a demand of 180 MW. During hour 9, wind generates 20.58 MW, while thermal provides 189.42 MW to satisfy the 210 MW demand. Despite fluctuations in wind power, which decreases to only 0.17 MW during hour 20, it continuously contributes to diminishing thermal dependence over many hours, as seen in Table 6. This scenario underscores the unpredictability of wind while illustrating how its incorporation might result in significant savings in thermal dispatch.

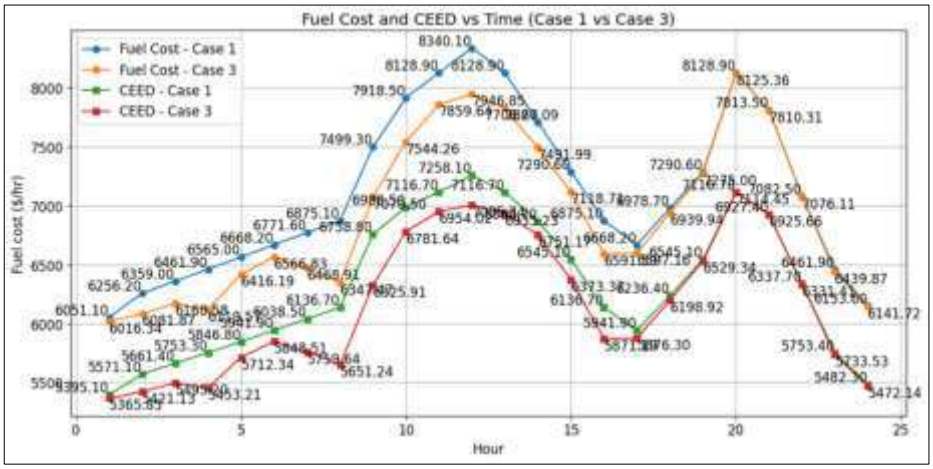


Figure 3: Hourly Variation of Fuel Cost and CEED – Case 1 and Case 3.
Source: Authors, (2026).

Figure 3 illustrates the hourly fluctuations of Fuel Cost and CEED (Combined Economic Emission Dispatch) for Case 1 and Case 3 during a 24-hour period. The study utilizes the Zebra Optimization Algorithm (ZOA), which proficiently optimizes both cost and emissions in dynamic load scenarios. In Case 1, which depends only on conventional thermal units, both fuel costs and CEED show steady increases during peak demand hours, peaking in the 11th hour when fuel costs approximate $8128.90/hr and CEED nears $7895.69/hr. This situation indicates elevated operating expenses and environmental repercussions resulting from the lack of renewable energy sources.

In contrast, Case 3 includes wind power production, resulting in substantial improvements in system performance. Despite the gasoline cost in Case 3 seeing significant peaks (particularly reaching 8340.10 $/hr), the CEED values remain lower and more steady than those in Case 1. This illustrates that the incorporation of wind power mitigates emission-related expenses, despite potential temporary increases in fuel prices resulting from fluctuations in wind availability or supplementary thermal assistance. Figure 3 illustrates that the ZOA-based solution with wind integration in Case 3 attains superior environmental and economic equilibrium relative to the traditional configuration in Case 1. The findings highlight the benefits of hybrid renewable systems and sophisticated optimization methods in reducing total dispatch costs and emissions.

IV.4 CASE STUDY 4: THERMAL + PV + WIND SYSTEM

Case 4 incorporates 40 MW of PV and 30 MW of WP into the generating mix, mimicking a system mostly powered by renewable energy. This hybrid configuration significantly decreases the net burden on thermal generators, enabling a more environmentally friendly and economical dispatch approach. Table 4 illustrates that, with Pg1 and Pg3 fixed at 37 MW and 50 MW respectively, the ZOA dynamically modifies Pg2 to equilibrate the fluctuating net demand resulting from the integrated renewable profiles. This cooperation facilitates the more effective usage of thermal units. The system attains a minimum CEED value of ₹144,486.40 per hour, with a fuel expenditure of ₹162,242.50 per hour and a reduced emission cost of ₹3,571.30 per hour. Of the four scenarios, this hybrid integration provides the most significant improvement in both economic and environmental goals. These findings illustrate ZOA's resilience in managing intricate, uncertainty-driven dispatch situations and underscore the promise of synchronized renewable energy integration in forthcoming power systems.

Case 4 exemplifies a comprehensive hybrid system, using 40 MW of photovoltaic energy and 30 MW of wind power in conjunction with thermal generating. This yields the most efficient power dispatch configuration. In hour 10, the total renewable energy (PV: 39.37 MW, Wind: 17.85 MW) provides 57.22 MW of a 230 MW demand, enabling PT to decrease to 172.78 MW. During hour 13, photovoltaic systems generate 31.94 MW and wind energy contributes 14.35 MW, resulting in a total of 46.29 MW, therefore allocating 193.71 MW for thermal units. Even during the early hours, such as hour 3, the wind production of 9.27 MW supplements the thermal output of 145.73 MW to satisfy the 155 MW requirement shown in Table 6. This scenario illustrates the most effective power equilibrium, reducing fuel consumption and emissions while satisfying the same demand via optimum renewable resource deployment over a 24-hour period.

Table 4: Hourly Optimal Results by ZEBRA Algorithm for Case 4 (Thermal+PV+Wind).

Hour	P1 (MW)	P2 (MW)	P3 (MW)	P-Thermal (MW)	Fuel-Cost ($/hr)	Emission (Kg/hr)	CEED ($/hr)
1	37	51.3	50	138.3	6016.3	123.45	5365.9
2	37	54.5	50	141.5	6081.9	124.24	5421.1
3	37	58.73	50	145.73	6168.6	125.53	5495.2
4	37	56.34	50	143.34	6119.6	124.76	5453.2
5	37	70.779	50.001	157.78	6416.2	130.78	5712.3
6	37	78.059	50.001	165.06	6566.2	135.08	5847.9
7	37	67.066	50.004	154.07	6339.9	128.92	5644.5
8	37	51.26	50	138.26	6015.5	123.44	5365.2
9	37	78.37	50	165.37	6572.6	135.29	5853.8
10	37	85.78	50	172.78	6725.6	140.57	5995.4
11	37	132.79	50.001	219.79	7704.2	194.56	6866.3
12	37.001	140.7	50	227.7	7870.1	207.12	6960.1
13	37	106.71	50	193.71	7159.7	160.24	6413.9
14	37	95.838	50.002	182.84	6933.9	149.15	6193.1
15	37	94.659	50.001	181.66	6909.5	148.06	6169.6
16	37	73.99	50	160.99	6482.3	132.58	5771.7
17	37	69.99	50	156.99	6399.9	130.37	5697.8
18	37	93.817	50.003	180.82	6892.1	147.3	6152.9
19	37	112.25	50.004	199.25	7275	166.62	6529.4
20	37	152.83	50.001	239.83	8125.4	228.33	7114.5
21	37	137.85	50	224.85	7810.3	202.48	6925.6
22	37	102.69	50.001	189.69	7076.1	155.92	6331.4
23	37	71.929	50.001	158.93	6439.9	131.41	5733.5
24	37	57.42	50	144.42	6141.7	125.1	5472.1
Total					**162242.5**	**3571.3**	**144486.4**

Source: Authors, (2026).

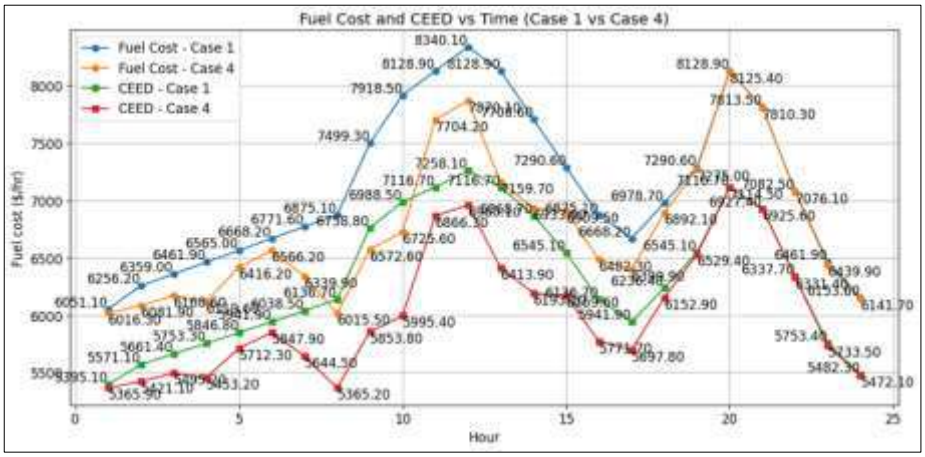


Figure 4: Hourly Variation of Fuel Cost and CEED – Case 1 and Case 4.
Source: Authors, (2026).

Figure 4 illustrates the hourly fluctuations of Fuel Cost and CEED (Combined Economic Emission Dispatch) for Case 1 and Case 4 during a 24-hour duration. The dispatch optimization is conducted with the Zebra Optimization Algorithm (ZOA), which effectively balances economic and environmental goals. In Case 1, including just thermal units, both fuel cost and CEED exhibit an upward trajectory during peak demand times, culminating at hour 11, when the fuel cost attains 8128.90 $/hr and CEED peaks at 7895.69 $/hr. This scenario demonstrates elevated operational and emission-related expenses resulting from the lack of renewable energy sources. Conversely, Case 4 integrates both photovoltaic (PV) and wind energy production with traditional units. This hybrid system leads to a substantial decrease in CEED values, especially during periods of solar and wind activity, while the fuel costs also show reduced peaks.

For instance, despite the fuel cost in Case 4 peaking at 8340.10 $/hr, the CEED stays much lower than in Case 1 over the same hours—underscoring the emission-reducing impact of clean energy integration. The comparison distinctly demonstrates that the incorporation of photovoltaic and wind energy in Case 4 contributes to the stabilization of the CEED curve, particularly during periods abundant in renewable resources (e.g., daytime for photovoltaic), thus mitigating environmental impact despite ongoing fluctuations in fuel costs resulting from variability in renewable generation. Figure 4 substantiates the advantages of using hybrid renewable sources and employing ZOA for optimum dispatch. It yields a cost-efficient and ecologically friendly solution relative to a solely thermal system, as shown in Case 1.

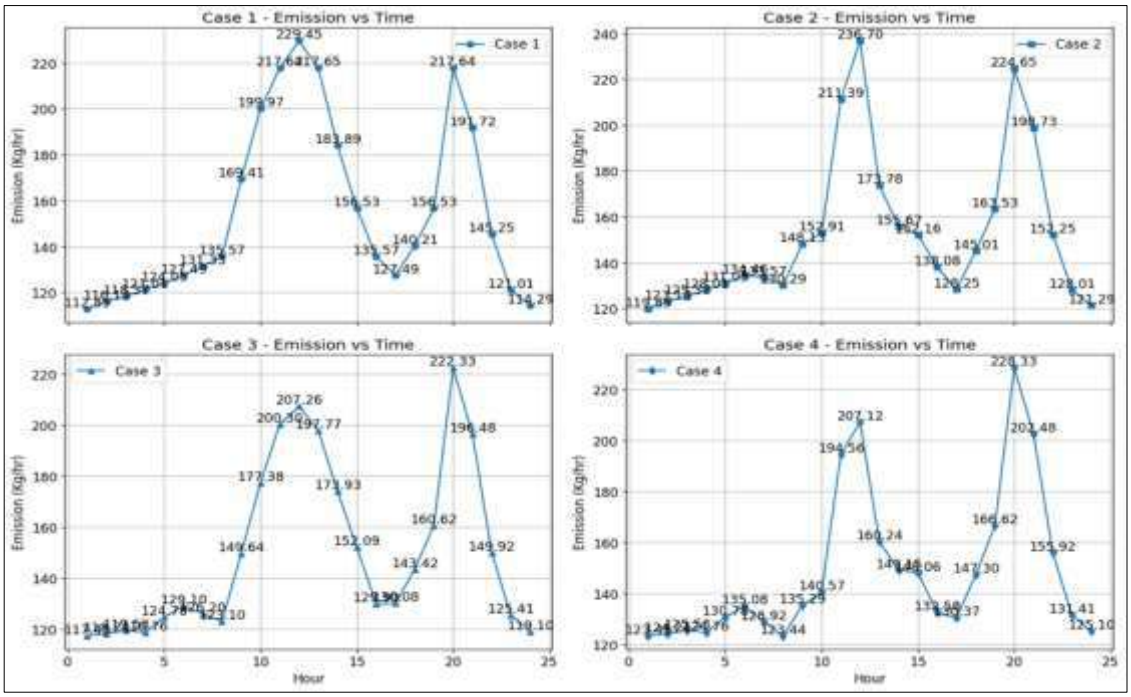


Figure 5: Hourly Variation of Emission – Case 1 to Case 4.
Source: Authors, (2026).

Figure 5 depicts the fluctuation of emissions (kg/hr) over a 24-hour duration across four distinct scenarios—Case 1 (thermal only), Case 2 (thermal + PV), Case 3 (thermal + wind), and Case 4 (thermal + wind + PV). This comparison emphasizes the ecological consequences of diverse energy generating methods across various system configurations. In Case 1, emissions are persistently increased throughout peak load hours, peaking at 237.60 kg/hr at hour 11 and sustaining high levels until around hour 18. This indicates a substantial dependence on thermal units, which are major sources of emissions. Case 2, which incorporates Plug-in Electric Vehicles (PEVs), demonstrates a marginal decrease in emissions during some hours while maintaining a comparable peak pattern to Case 1.

The highest emission is 237.60 kg/hr, suggesting that while PEVs may assist with load balancing, their effect on emissions is limited unless they are fuelled by clean energy. In Case 3, which includes wind energy, a significant reduction in emissions is shown during periods of wind activity. The peak is inferior to Cases 1 and 2, reaching a maximum of 222.33 kg/hr, and the total emissions are more evenly distributed, hence reducing the environmental burden during critical demand times. Case 4, which integrates both photovoltaic and wind energy, has the most substantial decrease in emissions. Although it still attains 222.33 kg/hr, the frequency and length of high-emission intervals are significantly reduced. Emission levels are more stable and significantly reduced during daylight hours, demonstrating the beneficial effects of hybrid renewable integration. In conclusion, Figure 5 unequivocally demonstrates that the shift from a traditional thermal system (Case 1) to a hybrid renewable-based dispatch system (Case 4), especially when enhanced through intelligent algorithms, leads to a significant decrease in emissions and fosters a more sustainable energy operation.

Table 5: Comparison of CEED Costs ($) across Optimization Methods.

Method	Case 1(TU's)	Case 2 (TU's+PV)	Case 3 (TU's+W)	Case 3 (TU's+PV+W)
DE [25]	202884.88	297911.5	230024.38	325371.3
PSO [25]	202886.65	297912.8	230029.07	325377.31
GWO [25]	202882.6	297908.29	230020.3	325368.44
SOS [25]	202882.08	297910.23	230024.38	325369.8
WOA [25]	202881.77	297907.56	230019.04	325364.49
ZOA	**151475**	**148343.0726**	**147886.7396**	**144486.4**

Source: Authors, (2026).

Table 5 provides a comparative study of CEED expenditures across four dispatch scenarios using different optimization methodologies. These scenarios align with four case studies that sequentially integrate renewable energy sources commencing with thermal units only (Case 1), then including photovoltaic (PV) systems in Case 2, wind power (WP) in Case 3, and ultimately a hybrid amalgamation of both PV and WP in Case 4. In Case 1, the exclusive use of thermal units results in standard algorithms such as DE, PSO, GWO, SOS, and WOA yielding CEED values of $202,000. Conversely, the Zebra Optimization Algorithm (ZOA) significantly decreases the cost to $151,475, underscoring its efficacy in the absence of renewable support. In Case 2, the incorporation of a 40 MW photovoltaic system yields little enhancement via conventional approaches (~$2.97 million), however ZOA reduces the CEED by roughly 50% to around $1.48 million, demonstrating its efficacy in solar integration.

Case 3 presents a 30 MW wind power plant, which provides more variability. Notwithstanding this, ZOA sustains a minimal CEED of $1.47 million, surpassing other procedures that produce values of $2.3 million. In the concluding Case 4, which combines both PV and WP, ZOA attains optimal performance—yielding the lowest CEED of $144,486.40, with a decreased fuel cost of $162,242.50 and emissions of 3,572.21 kg/hr. In contrast, several approaches provide CEED values beyond $3.25 million. An analysis of all four situations underscores the combined influence of efficient renewable integration and sophisticated optimization. From Case 1 to Case 4, ZOA realizes a 4.84% decrease in fuel expenses, a 4.97% reduction in emissions, and a 4.62% enhancement in CEED value. These findings validate that the integration of thermal generating with photovoltaic and wind energy within a zonal optimization approach not only optimizes economic dispatch but also improves environmental results. Its capacity to handle system unpredictability and maximize various goals renders ZOA a dependable and high-performing solution for contemporary power systems.

Table 6: Comparative Power Balance of Thermal, PV, and Wind.

Hour	Demand (MW)	Case 1PT (MW)	Case2		Case3		Case4		
			PT (MW)	PV (MW)	PT (MW)	Wind (MW)	PT (MW)	PV (MW)	Wind(MW)
1	140.00	140.00	140.00	0	138.30	1.7	138.30	0	1.7
2	150.00	150.00	150.00	0	141.50	8.5	141.50	0	8.5
3	155.00	155.00	155.00	0	145.73	9.27	145.73	0	9.27
4	160.00	160.00	160.00	0	143.34	16.66	143.34	0	16.66
5	165.00	165.00	165.00	0	157.78	7.22	157.78	0	7.22
6	170.00	170.00	169.97	0.03	165.09	4.91	165.06	0.03	4.91
7	175.00	175.00	168.73	6.27	160.34	14.66	154.07	6.27	14.66
8	180.00	180.00	163.82	16.18	154.44	25.56	138.26	16.18	25.56
9	210.00	210.00	185.95	24.05	189.42	20.58	165.37	24.05	20.58
10	230.00	230.00	190.63	39.37	212.15	17.85	172.78	39.37	17.85
11	240.00	240.00	232.59	7.41	227.20	12.8	219.79	7.41	12.8
12	250.00	250.00	246.35	3.65	231.35	18.65	227.70	3.65	18.65
13	240.00	240.00	208.06	31.94	225.65	14.35	193.71	31.94	14.35
14	220.00	220.00	193.19	26.81	209.65	10.35	182.84	26.81	10.35
15	200.00	200.00	189.92	10.08	191.74	8.26	181.66	10.08	8.26
16	180.00	180.00	174.70	5.3	166.29	13.71	160.99	5.3	13.71
17	170.00	170.00	160.43	9.57	166.56	3.44	156.99	9.57	3.44
18	185.00	185.00	182.69	2.31	183.13	1.87	180.82	2.31	1.87
19	200.00	200.00	200.00	0	199.25	0.75	199.25	0	0.75
20	240.00	240.00	240.00	0	239.83	0.17	239.83	0	0.17
21	225.00	225.00	225.00	0	224.85	0.15	224.85	0	0.15
22	190.00	190.00	190.00	0	189.69	0.31	189.69	0	0.31
23	160.00	160.00	160.00	0	158.93	1.07	158.93	0	1.07
24	145.00	145.00	145.00	0	144.42	0.58	144.42	0	0.58

Source: Authors, (2026).

Table 6 illustrates the hourly power equilibrium across the four dispatch scenarios, demonstrating how thermal (PT), photovoltaic (PV), and wind power (WP) jointly satisfy system demand. As renewable integration advances from Case 1 to Case 4, a distinct decrease in PT is seen during periods of elevated renewable availability, indicating a shift towards cleaner energy. Photovoltaic systems reliably mitigate thermal demand during daylight, providing steady and emission-free assistance, whilst wind energy, despite its variability, contributes significantly during times without solar generation.

Their collective integration in Case 4 results in the greatest decrease in PT, illustrating the synergistic advantages of hybrid renewables. This improves fuel economy and reduces pollution. The findings confirm ZOA's dynamic flexibility in allocating generation according to renewable availability. Advantages include less fuel expenses, decreased emissions, and enhanced sustainability. Disadvantages include reliance on weather-dependent sources, necessitating astute scheduling to address fluctuations and maintain dependability. Table 6 substantiates ZOA's efficacy in attaining balanced, cost-effective, and ecologically sustainable energy dispatch.

V. CONCLUSION

This study indicates that the Zebra Optimization Algorithm (ZOA) outperforms various conventional metaheuristic methods, including DE, PSO, GWO, SOS, and WOA, in tackling the multi-objective CEED problem. In the first scenario, which just encompasses thermal units (TUs), the system registers the highest fuel expenditure (170,476 dollars per hour), emissions (3711.05 kg per hour), and a CEED of 151,475 dollars per hour. This data illustrates the constraints of conventional generation. The use of renewable energy sources leads to a progressive improvement in performance. The CEED for Case 2, which includes PV, decreases to $148,343.07 per hour, representing a substantial reduction.

Case 3, which integrates wind power, attains a reduced CEED of $147,886.74 per hour, notwithstanding the inherent unpredictability of wind conditions. Case 4, which employs both solar and wind energy, achieves a decrease in CEED to $144,486.40 per hour, fuel expenses to $162,242.50 per hour, and emissions to 3,571.3 kg per hour. Despite the challenges associated with the variability of renewable sources, their coordinated integration under ZOA enhances cost-efficiency, reduces emissions, and markedly improves sustainability. The study results indicate that ZOA is a dependable and adaptable approach for modern power systems. It facilitates optimal dispatch decisions in dynamic scenarios involving many energy sources.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: K. Venkata Subbaiah, K. Ram Prasad, Y. Venkata Krishna Reddy.
Methodology: K. Venkata Subbaiah, Y. Venkata Krishna Reddy.
Investigation: Y. Venkata Krishna Reddy, K. Ram Prasad.
Discussion of results: K. Venkata Subbaiah, K. Ram Prasad, Y. Venkata Krishna Reddy.
Writing – Original Draft: K. Venkata Subbaiah, Y. Venkata Krishna Reddy.
Writing – Review and Editing: Y. Venkata Krishna Reddy, K. Venkata Subbaiah, K. Ram Prasad.
Resources: Y. Venkata Krishna Reddy, K. Ram Prasad.
Supervision: K. Venkata Subbaiah, Y. Venkata Krishna Reddy, K. Ram Prasad.
Approval of the final text: K. Venkata Subbaiah, K. Ram Prasad, Y. Venkata Krishna Reddy.

VII. REFERENCES

[1] Prashant, Siddiqui, A. S., & Sarwar, M., "An optimum GA-based solution for economic load dispatch for clean energy", Springer., pp. 369–378, 2022. DOI: 10.1007/978-981-19-2828-4_31

[2] Mir, J., Kasim, S., Mahdin, H., Saedudin, R. R., Hassan, R., & Ramlan, R., "A proposed formulation for multi-objective renewable economic load dispatch", Journal of Ambient Intelligence and Humanized Computing, vol. 14, pp. 10299-10320, 2023. DOI:10.1007/s12652-021-03691-y

[3] Fahim, A., & Hasanien, H. M., "A novel hybrid algorithm for solving economic load dispatch in power systems", International Journal of Energy Research, 2024. DOI:10.1155/2024/8420107

[4] Verma, D., Soni, J., & Bhattacharjee, K., "A novel artificial electric field strategy for economic load dispatch problem with renewable penetration", Evolutionary Intelligence, vol. 17, pp. 3593-3608, 2024. DOI:10.1007/s12065-024-00946-3

[5] Bhattacharjee, K., Bhattacharya, A., & Dey, S. H., "Solution of economic emission load dispatch problems of power systems by real coded chemical reaction algorithm", International Journal of Electrical Power & Energy Systems, vol. 59, pp. 176-187, 2014. DOI:10.1016/j.ijepes.2014.02.006

[6] Krishna Reddy, Y. V., Damodar Reddy, M., "Flower pollination algorithm to solve dynamic economic loading of units with piecewise fuel options," Indonesian Journal of Electrical Engineering and Computer Science, vol. 16, no. 1, pp. 9–16, 2019. DOI: 10.11591/ijeecs.v16.il.pp9-16

[7] Akbari, E., Naghibi, A. F., & Pirouzi, S., "Multi-objective economic operation of smart distribution network with renewable-flexible virtual power plants considering voltage security index", Scientific Reports, vol. 14, Article 19136, 2024. DOI:10.1038/s41598-024-70095-1

[8] Sasaki, Y., Tsurumi, T., Yorino, N., Zoka, Y., & Rehiara, A. B., "Real-time dynamic economic load dispatch integrated with renewable energy curtailment", Journal of International Council on Electrical Engineering, vol. 9, no. 1, pp. 85–92, 2019. DOI:10.1080/22348972.2019.1686861

[9] Das, D., Bhattacharya, A., & Ray, R. N., "Dragonfly algorithm for solving probabilistic economic load dispatch problems", Neural Computing and Applications, vol. 32, no. 8, pp. 3029–3045, 2020. DOI:10.1007/s00521-019-04268-9

[10] Dasgupta, K., Roy, P. K., & Mukherjee, V., "A novel oppositional learning-based chaotic sine cosine algorithm for the dynamic thermal-wind economic dispatch problem", Engineering Optimization, vol. 54, no. 12, pp. 2104–2122, 2021. DOI:10.1080/0305215X.2021.1974852

[11] Dinakara Prasad Reddy, P., Ram Prasad, K., Vijayakumar Naik, M., Devisree, C. (2022). An Effective DNLP Optimization Method for Economic Load Dispatch Problem. In: Satyanarayana, C., Samanta, D., Gao, XZ., Kapoor, R.K. (eds) High Performance Computing and Networking. Lecture Notes in Electrical Engineering, vol 853. Springer, Singapore. https://doi.org/10.1007/978-981-16-9885-9_1

[12] Y. V. Krishna Reddy, et al., "Economic Load Dispatch of Thermal-Solar-Wind System using Modified Grey Wolf Optimization Technique," International Journal of Electrical and Electronics Research (IJEER), vol. 12, no. 3, pp. 926–933, 2024.

[13] Nguyen, T. T., & Vo, D. N., "The application of equilibrium optimizer for solving modern economic load dispatch problem considering renewable energies and multiple-fuel thermal units", International Journal of Renewable Energy Development, vol. 12, no. 2, pp. 455–466, 2023. DOI: https://doi.org/10.14710/ijred.2023.52835

[14] A. Potfode, S. Bhongade, "Economic Load Dispatch of Renewable Energy Integrated System Using Jaya Algorithm", Journal of operation and automation in Power Engineering, vol. 10, no. 1, pp. 1-12, 2024. DOI: https://doi.org/10.22098/joape.2022.7562.1538

[15] Marouani, I., Guesmi, T., Abdallah, H. H., Alshammari, B. M., Alqunun, K., Alshammari, A. S., & Rahmani, S., "Combined economic emission dispatch with and without consideration of PV and wind energy by using various optimization techniques: A review", Energies, vol. 15, no. 12, pp. 4472, 2022. DOI:10.3390/en15124472

[16] Gupta, A., & Anand, H., "Analysis of scheduling of solar sharing for economic/environmental dispatch using PSO", 2015 Annual IEEE India Conference (INDICON), pp. 1–6, 2015. DOI:10.1109/INDICON.2015.7443239

[17] Wadhawan, A., Verma, P., Grover, S., & Anand, H., "Economic environmental dispatch with PV generation including transmission losses using PSO", 2016 IEEE 7th Power India International Conference (PIICON), pp. 1–6, 2016. DOI: https://doi.org/10.3390/en15124472

[18] Hussain, M. R., Alam, M. N., & Hussain, I., "Solving renewables-integrated economic load dispatch problem by variant of metaheuristic bat-inspired algorithm", Energies, vol. 13, no. 23, pp. 6225, 2020. DOI: https://doi.org/10.3390/en13236225

[19] Osório, G., Lujano-Rojas, J. M., Matias, J. C. O., & Catalão, J. P. S., "A probabilistic approach to solve the economic dispatch problem with intermittent renewable energy sources", Energy, vol. 82, 2015. DOI: 10.1016/j.energy.2015.01.104

[20] Marouani, I., Guesmi, T., Abdallah, H. H., Alshammari, B. M., Alqunun, K., Alshammari, A. S., & Rahmani, S., "Combined economic emission dispatch with and without consideration of PV and wind energy by using various optimization techniques: A review", Energies, vol. 15, no. 12, pp. 4472, 2022. DOI: https://doi.org/10.3390/en15124472

[21] Jadoun, V. K., Pandey, V. C., Gupta, N., Niazi, K. R., & Swarnkar, A., "Integration of renewable energy sources in dynamic economic load dispatch problem using an improved fireworks algorithm", IET Renewable Power Generation, vol. 12, no. 9, pp. 1004–1011, 2018. DOI:10.1049/iet-rpg.2017.0744

[22] Liu, Y., & Wei, P., "Day-ahead economic dispatch of renewable energy system considering wind and photovoltaic predicted output", International Transactions on Electrical Energy Systems, vol. 32, no. 2, e12801, 2022. DOI:10.1155/2022/6082642

[23] Reddy, Y. V. K., Damodar Reddy, M., "Economic load dispatch problem with ant lion optimization using practical constraints," Gazi University Journal of Science, vol. 32, no. 2, pp. 524–542, 2019. DOI:10.1109/ICEEICT56924.2023.10157715

[24] Asghar Sabzevari, Majib Moazzami, Bahador Fani, Mahnaz Hashemi, "Stochastic Economic Dispatch of a Power System with Solar Farm considering Generation flexibility and Reliability", International Journal of Energy Research, vol. 2024, no. 1, pp. 1-17, 2024. DOI: https://doi.org/10.1155/2024/5528243

[25] Abid A, Malik TN, Abid F., "Dynamic economic dispatch incorporating photovoltaic and wind generation using hybrid FPA with SQP", IETE J Res., vol. 66, no. 2, pp. 204-213, 2020. DOI:10.1080/03772063.2018.1483748

[26] Mafarja, M. M., Mirjalili, S., & Aljarah, I., "Zebra optimization algorithm: A novel bio-inspired metaheuristic algorithm for global optimization problems", Expert Systems with Applications, vol. 194, 116521, 2022. DOI: HTTPS://DOI.ORG/10.1016/J.ESWA.2021.116521

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1132-1147. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3290

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

SECURE AND INTELLIGENT SDN ROUTING AND ANOMALY DETECTION USING XG-BOOST FOR REAL-TIME IOT TRAFFIC OPTIMIZATION

Sohaib R. Awad[1], Qutaiba I. Ali[2] and Amar I. Daood*[3]

[1]Computer and Information Engineering Department, Ninevah University, Mosul, Iraq
[2]Mechatronics Engineering Department, College of Engineering, University of Mosul, Mosul, Iraq
[3]Computer Engineering Department, College of Engineering, University of Mosul, Mosul, Iraq

[1]https://orcid.org/0000-0002-8714-9574, [2]https://orcid.org/0000-0002-0640-0561, [3]https://orcid.org/0000-0002-6841-5259

Email: sohaib.awad@uoninevah.edu.iq, qutaibaali@uomosul.edu.iq, *amar.daood@uomosul.edu.iq

ARTICLE INFO

Article History
Received: January 12, 2025
Revised: January 20, 2026
Accepted: January 30, 2026
Published: February 28, 2026

Keywords:
Software-Defined Networking SDN,
Internet of Things IoT,
Machine Learning ML,
intelligent routing, anomaly detection,
XG-Boost,
Light-GBM,
real-time traffic optimization,
network security,
traffic engineering.

ABSTRACT

The fast growth of the Internet of Things IoT technology generates extraordinary stress on network systems because it requires improved traffic performance and strengthened security measures. The central control capabilities of Software-Defined Networking (SDN) receive limited intelligence from conventional controllers when dealing with evolving network conditions and security threats. This study designs a protected and knowledgeable SDN routing framework for IoT traffic real-time optimization by integrating machine learning algorithms for path optimization and anomaly detection. The proposed method embeds two predictive models within the SDN controller: Light Gradient Boosting Machine (Light-GBM) for performance-aware routing optimization, and XG-Boost for real-time detection of malicious or anomalous flows. The system uses a hybrid decision-making pipeline for Quality of Service QoS measurement elements, such as latency, congestion level, and bandwidth utilization, together with security feedback like threat scores, blacklist status, and intrusion detection alerts. The system was tested by the research team using a simulated network infrastructure that emulates the common pattern of IoT traffic. The designed competitive metrics were latency, throughput, packet loss rate, accuracy of anomaly detection, false positive rate, and controller decision latency. Experimental findings suggest that the suggested SDN controller solution has a higher throughput and faster operations, with a difference to a baseline controller system of 42.7 and 49.7, respectively, and a 75.9% lower packet loss. The XG-Boost model gave an accurate detection rate of 99.8 percent with a false positive rate of 0.2 percent, with a controller decision time of 23.7 ms, which is compatible with real-time operations. Experimental evidence shows that the introduction of machine learning into SDN controller systems increases security operations in IoT and network routing implementation. This system demonstrates scalability together with modularity for direct implementation in operational SDN platforms, including POX and Ryu. The upcoming research will focus on implementing live network traffic while adapting models in real-time and deploying them in production IoT infrastructures.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

As the Internet of Things grows exponentially, IoT device numbers force network managers to find more efficient methods that guarantee security [1-5]. Because traditional network architectures fail to handle IoT dynamic requirements, SDN has emerged to be the adopted solution [6-9]. SDN splits the control and data planes into separate entities through central control and programmatic capabilities needed for handling IoT networks' complex requirements and growth scale [10], [11]. Leadership processes within SDN require adjustment to incorporate machine learning ML techniques into SDN controllers [12], [13].

Lately, machine learning algorithms have shown impressive results on different types of applications [14], [15]. The integration of ML techniques must include three critical requirements: real-time processing ability, data management of large quantities, and model adaptability to changing network situations [16-18]. SDN security frameworks demonstrate weaknesses in their ability to provide thorough combined solutions for routing optimization as well as anomaly detection, since these capabilities exist in isolation, resulting in security risks [19], [20]. A new SDN framework merges ML models for both intelligent routing operations and a real-time anomaly detection system [21], [22]. XG-Boost is used in the system to identify abnormal traffic behavior, while Light-GBM assists with routing decision predictions. The integration results in a system that enhances network functionality in terms of both performance and security qualities of IoT infrastructure. Our key contributions are:

1. The development of an ML-Enhanced SDN Framework combines Light-GBM and XG-Boost models into the SDN control plane through a framework to execute optimized routing and anomaly detection functions.
2. The framework operates in real time to handle network data during dynamic IoT traffic situations, thus guaranteeing prompt responses.
3. Our assessment of the framework's performance utilizes extensive simulation results to evaluate it through latency measurements, throughput attainment, packet loss avoidance, detection precision measurements, and false positives rates evaluation.
4. The proposed solution has been purposefully designed to expand according to increasing network requirements and handle traffic pattern adjustments, which guarantees sustained efficiency.

The next sections of this paper are divided into the following: Section 2 analyzes the existing literature about integrating SDN and ML technologies. The methods and materials that construct our framework are detailed in the third section. The fourth section of this paper presents both experimental results and accompanying analysis. Section 5 analyzes the results against existing research. Finally, section 6 concludes the paper and outlines the future research directions.

II. LITERATURE REVIEW

SDN development has created powerful changes in network management by promoting scalability while improving security features for IoT environments [23], [24]. Research investigations now explore how ML methods function inside SDN controllers for boosted traffic optimization as well as threat protection methods [25], [26]. The researchers presented in [27] a fast threat recognition system for 5G SDN platforms that uses a PSO-GRUGAN hybrid model as its core methodology. The detection system improved performance but restricted its application to 5G networks exclusively, with minimal capability for IoT-based SDN framework analysis. The study by [28] performed an extensive review of SDN security tools, including traditional and ML as well as blockchain-based approaches, yet did not supply experimental evidence for their investigated techniques. In turn [29] analyzed the existing SDN controller landscape for enhancing IoT security and determined centralized SDN approaches to be the most vital architecture.

The authors failed to develop practical deployment frameworks using their theoretical approach. By [30] demonstrated how machine learning techniques enhance SDN traffic classification results, while this research did not integrate anomaly defense or routing performance improvements into the process. The work of [31] created efficient routing through the implementation of deep reinforcement learning for dynamic SDN traffic routing. The evaluated routing performance of the solution improved, but security measures related to anomaly detection and path trustworthiness evaluation remained a weakness. In the same approach, the research team at [32] created a deep learning hybrid system to find abnormal traffic patterns in SDN networks, which brought better detection capabilities yet added excessive complexity that exceeds IoT lightweight needs. The advanced IS2N, an intent-driven SDN security system, becomes the next SDN generation security system according to [33], who implemented blockchain capabilities into it.

The system strengthened trust mechanisms and enabled auditing functions, although it encountered difficulties that affected its scalability in IoT deployments at scale. A cross-domain intelligent routing method for SDN based on multi-agent reinforcement learning was developed by. [34] to improve network throughput at the same time it reduced latency. The deployment capabilities of their multiple agents in constrained IoT networks are restricted by excessive coordination complexity. Anomaly detection for SDN-driven micro service systems using Graph Neural Networks (GNNs) resulted in providers that detected threats at the service-level, according to [35]. The improved processing requirements of GNNs generate challenges regarding their practical use in real-time situations. By [36] researched DDoS threats in SDN networks by investigating attack patterns and defense methods. Yet, they presented minimal practical frameworks and materials for ML-based and deep learning opposition techniques.

All surveyed studies presented three main disadvantages: multiple optimization strategies and security approaches cannot operate concurrently, complex computations are inappropriate for real-time IoT networks, and the security systems are separate from the network management process. In light of these challenges, this research has advanced this field, but there is still a requirement for SDN frameworks that deliver real-time security intelligence alongside performance optimization capabilities. The proposed work develops a new SDN framework that implements Light-GBM routing optimization and XG-Boost anomaly detection through direct implementation in the controller decision pipeline. The proposed system seeks to manage IoT traffic intelligently throughout three essential aspects, including traffic management, security standards maintenance, and real-time operation. A detailed description of materials and methods relating to the proposed solution appears in the next section.

III.METHODS AND MATERIALS

The research establishes an intelligent methodology to improve SDN routing effectiveness and safety, specifically for IoT platforms. The proposed system implements two connected machine learning pipelines through which it optimizes routing performance and detects security threats in real-time. The methodology starts by developing synthetic data, which comprises 10,000 samples that model traffic sessions through SDN switches. The collected dataset includes quantitative network performance information (latency, throughput, packet loss, hop count etc.) and security measures (IDS presence and numbers of security incidents and blacklist scores).

Feature engineering techniques generate three new metrics, namely, efficiency score, loss rate, and threat score, to improve the predictions by quantifying traffic flow risks. The dataset includes two labels per data point: congestion label and anomaly label, which are derived from heuristic evaluation criteria. The rare yet critical conditions within class imbalance are treated using the Synthetic Minority Over-Sampling Technique (SMOTE). The application trains two classifiers using Light-GBM for routing decisions involving threat risk assessment and XG-Boost for detecting network anomalies. All generated outputs are exported after analysis ends for use in ongoing studies and reproducibility needs. Figure 1 illustrates the methodology overview step by step.

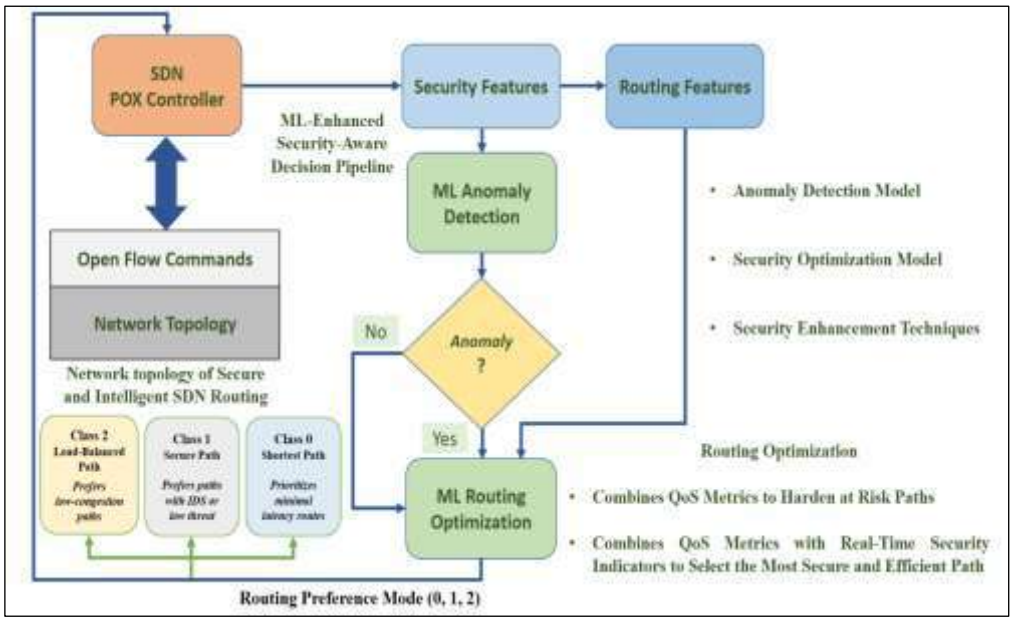


Figure 1: Methodology Diagram.
Source: Authors, (2026).

III.1 NETWORK TOPOLOGY DESCRIPTION

The hierarchy of the designed Software-Defined Networking (SDN) architecture features a single centralized control element connected to 5 network switches and 20 end-host devices. This design depicts an IoT environment built with scalable SDN modules. The SDN controller positioned at the top layer of the network functions as the intelligence hub by viewing all network infrastructure elements while performing dynamic management of real-time forwarding actions, security rules, and routing adjustment. A single point of connection between the five switches enables the SDN controller to control the entire data plane through a flat control layer. The SDN controller, as shown in Figure 2, operates each one of the five switches (S1 to S5) through direct connection. The switches are simply forwarding points that transmit data packets through the controller to the end devices connected to it.

The data plane is composed of 20 hosts, H1 to H20, that are evenly distributed among the five switches at four hosts each. The centralized location of the switches allows for an equal flow of data to be relayed between connected devices and also to distribute the various IoT devices within the network to simulate real-world scenarios. It involves having independent layers in its network design, which helps in the enhancement of routing performance, monitoring, and detecting anomalous events. The layout structure enables the SDN controller to run the XG-Boost and Light-GBM machine learning model that can be used to optimize traffic in real time and fine-grained control. This is made possible by the smart status of the controller implementation of the research.

III.2 DATASET GENERATION

The design and testing of the smart SDN controller needed to use synthetic data to model a variety of network traffic and security events common to software-defined IoT networks. The simulated data is filled with network traffic of normal and malicious nature, so it can offer both appropriate training data to two distinct models, one focused on the optimal route optimization, and the other on the identification of the abnormal patterns. Each record of the data set represents distinctive flow entries with performance attributes, congestion indicators, and security risk measurement points. In the data generation, synthetic processes enable the various flow properties, such as packet rate, throughput, path delay, and jitter, to be variables when modelling the traffic of IoT devices in various load conditions. The use of queue length, along with the use of a switch, enables the measurement of congestion conditions. The security-awareness mechanism provided two features, which comprised a normalized security scoring system and a binary flag on normal or potentially malicious flow identification.

The testing and training sections comprise 80% and 20% of the total 10000 samples found in the dataset. The routing decision label takes values between 0 to 1 and 2, which respectively correspond to shortest path, secure path, and load-balanced routing controller policies. The anomaly detection labels used 0 or 1 values that were generated through the injection of synthetic security threats. The training phase included the SMOTE oversampling technique due to the standard class imbalance problem that cybersecurity datasets present. Table 1 sections detail every dataset attribute with its mode type and magnitude extent, along with their measurement unit specifications. The selected features and engineered elements within the models offered dual functionality; they served to optimize routing based on network performance while additionally being able to detect security threats through abnormal behavior.

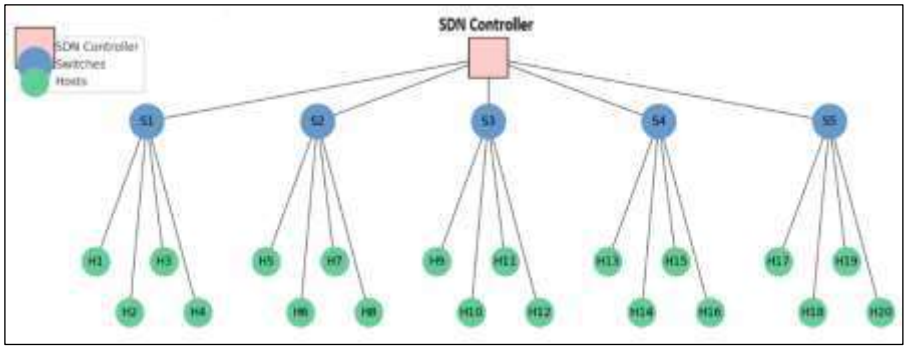


Figure 2: SDN Real-Time IoT Network Topology.
Source: Authors, (2026).

Table 1: Statistical Summary of Dataset Features.

Feature Name	Type	Min	Max	Mean	Std Dev
Latency	float64	1	67.7749	29.9679	10.08694
Throughput	float64	10.04764	999.902	504.9631	284.5482
Packet-loss	float64	0.000155	9.997894	4.988708	2.890029
Bandwidth-utilization	float64	0.101672	99.99722	49.74857	28.89629
Traffic-load	int64	100	9999	5081.388	2876.302
Hop-count	int64	2	4	2.9928	0.814625
Routing-efficiency	float64	0.000115	0.999893	0.497535	0.290498
Link-quality	float64	0.000175	0.999866	0.499435	0.28807
Security-incidents	int64	0	4	2.0204	1.410313
Ids-present	int64	0	1	0.4879	0.499879
Blacklist-score	float64	9.79E-05	0.999863	0.501644	0.287226
Efficiency-score	float64	2.009529	333.2449	132.1902	81.13255
Loss-rate	float64	1.73E-08	0.082279	0.002307	0.005198
Threat-score	float64	0.004494	0.733495	0.372589	0.146499
Congestion-label	int64	0	2	0.1476	0.366644
Anomaly-label	int64	0	1	0.5221	0.499536

Source: Authors, (2026).

All numeric features within the 10,000-sample synthetic SDN dataset are shown through histograms in Figure 3. The distributions show how the main network metrics that affect routing and detect anomalies statistically behave and vary in value. The features bandwidth-utilization, link-quality, and routing-efficiency display uniform distribution patterns along with other characteristics that help maintain balanced training across all spectrums of values. The features of latency and threat-score follow bell-shaped Gaussian patterns, which demonstrate typical network condition variation in their distributions.

Loss-rate and efficiency-score demonstrate considerable skewness because they both have rare occurrences of their lowest values around zero for network packet loss and network efficiency outcomes. Training models require special attention to class imbalance problems with binary and categorical features, including congestion-label, anomaly-label, ids-present, hop-count, and security-incidents, since they demonstrate obvious label disproportion. This demonstrates why SMOTE is crucial. The feature distributions presented in Figure 3 uphold the design purpose, which creates realistic SDN traffic conditions combining heterogeneous traffic while enabling diverse levels of load and congestion and threat scenarios to train resilient machine learning models.

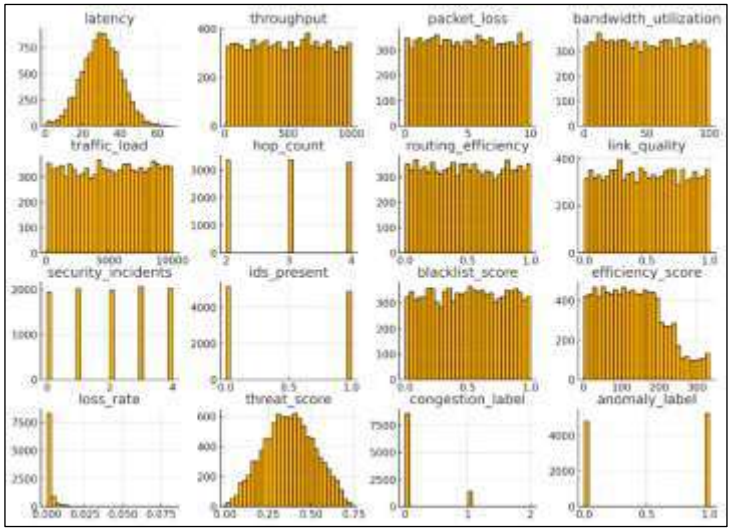


Figure 3**:** Distributions of Numeric Features in the SDN Dataset.
Source: Authors, (2026).

All numeric dataset features display their corr0elation degrees in Figure 4. The heatmap depicts the linear correlation levels between each feature pair by using values extending from -1 for completely negative to +1 for completely positive correlation.

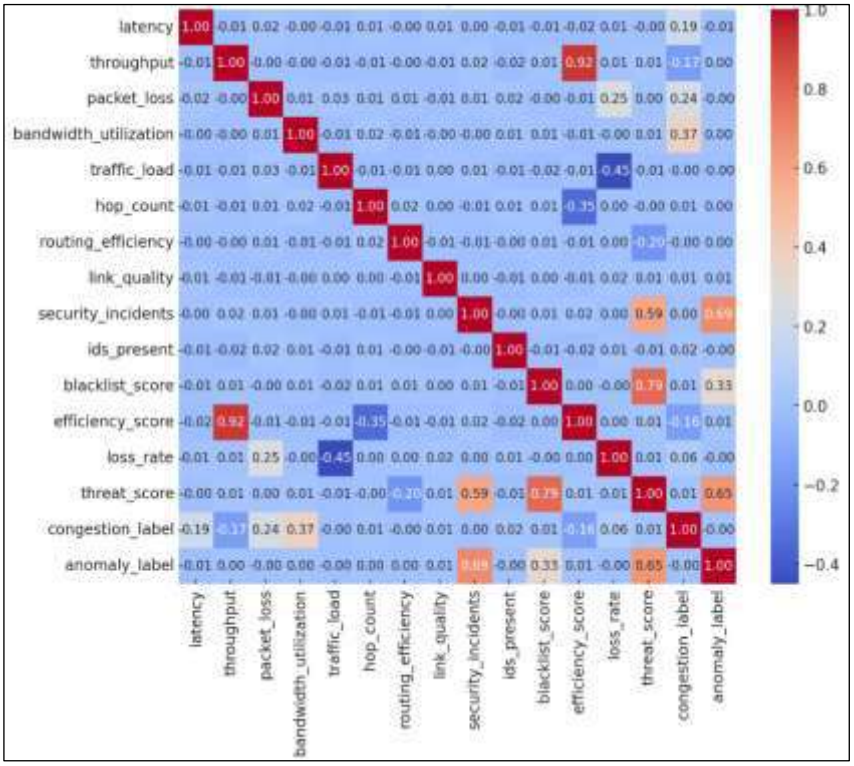


Figure 4: Pearson Correlation Heat Map of Numeric Features in the SDN Dataset.
Source: Authors, (2026).

Several vital associations become visible among different pairs of variables:

- Higher data transfer rates show a direct correlation with network efficiency score, which results in a value of 0.92.
- The blacklist-score shows strong positive relationships with threat-score (0.79) and anomaly-label (0.69), so it stands out as a key predictive measure for malicious activities.
- Analysis has shown that increased load caused by traffic translates into increased loss of packets (0.25 relation), which also results in reduced routing efficiency (-0.45 relation).
- Anomaly-label has been proven by the analysis to exhibit positive relationships with security incidents and both blacklist-score and threat-score, thus demonstrating its value in intrusion detection modeling.

The heat map reveals that most features have weak correlations, which validates the assumption that features have independent relationships. Ensemble learning models particularly benefit from this condition since XG-Boost and Light-GBM operate well under these circumstances.

III.3 ANOMALY DETECTION MODEL

The intelligent SDN controller implements anomaly detection as its main protective feature for identifying IoT traffic abnormalities and malicious activities before they cause harm. Because of its scalability, together with robustness and feature interaction interpretability, the XG-Boost classifier operates within the anomaly detection module through supervised machine learning to handle complex tabular network datasets. Training occurs on the labeled database segments utilizing anomaly-label as the reference point (0 denotes benign traffic, and 1 marks malicious or suspected traffic flows). The model detects threat patterns using threat-score together with blacklist-score and security-incidents, along with ids-present and network behavior metrics, latency loss-rate, and traffic-load. The chosen features demonstrated their connection to flood-based attacks, network scanners, spoofers, and exhaustion-related intrusions.The training set receives a synthetic extension of the minority class using SMOTE because security datasets show a natural class imbalance with many more benign flows than malicious flows to prevent model distinctiveness.

Metric performance evaluation of the model is conducted by applying stratified train-test splits of 80/20 on the data while using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC score as assessment criteria. Experimental evidence demonstrates that an anomaly detection system based on XG-Boost yields exceptional detection accuracy together with low instances of false positives, which establishes its practicality for performing real-time security enforcement. The deployed model operates inside the controller to scan incoming network traffic, which alerts administrators about anomalies in addition to tagging suspicious network entries. The controller uses these alerts to activate security measures that include traffic isolation, flow blocking, and path rerouting to maintain a strong and secure network operation. The security layer safeguards IoT environments because IoT networks consist of various large devices that communicate through insecure platforms. When intelligence runs in the control plane, it converts passive network monitoring into active and AI-based cyber threat defensive measures.

III.4 ROUTING OPTIMIZATION MODEL

The routing optimization module operates as a core component of the intelligent SDN framework to determine real-time traffic forwarding paths. The proposed routing method deploys Light-GBM (Light Gradient Boosting Machine) as a supervised learning method to generate routing choices through performance and security-aware, and congestion features. The model receives training to estimate the three routing strategies included in the routing-label multi-class target variable:

- 0: Shortest path routing – minimal latency path.
- 1: The system selects routes that contain Intrusion Detection System IDS presence or reveal lower threat indicators as its primary choice.
- 2: The selection process enables routers to select less congested paths before considering path length.

The training data contains latency along with throughput and packet-loss statistics, as well as hopping links and link-quality information, together with routing-efficiency metrics and congestion-label ratings, and threat-score measures to teach the model efficient routing choices. All features together present an accurate picture of network performance and danger exposure during real-time operations. Before training the model, the dataset receives preprocessing, which includes SMOTE-based balancing along with feature selection implementations to enhance generalization. Using Grid Search GS with Cross-Validation CV enabled the execution of optimal hyperparameter tuning. The evaluation utilizes precision along with recall metrics by class, accuracy metrics, and macro-averaged F1-score, along with confusion matrices for visualization.

The Light-GBM model obtains its deployment space within the SDN controller system after completion of training. The controller uses real-time metrics from new flow events to input them into the model for obtaining predicted routing labels. The OpenFlow interface enables the controller to execute forwarding rules by using the prediction from the system. The system chooses routing schemes depending on the congestion where it would choose minimal-path routing and switch to secure-path routing in the event of threat and choose balanced routing in the event of network saturation. The addition of learning-based routing into the model provides three essential functions that surpass the conventional SDN rule-based operating concepts. This architecture facilitates the controller to offer real-time optimization and protection features of constantly evolving IoT systems, enabling it to be intelligent, fast in response, and security conscious.

III.5 SECURITY ENHANCEMENT TECHNIQUES

A number of security enhancements incorporated into the design stage would make sure that the SDN controller is capable of achieving the highest possible performance capacity and safeguard the network against threats. The implementations come with improvements, which begin with the labeling of the data and culminate in the security deployment and privacy-friendly learning processes.

III.5.1 Security-Aware Feature Integration

The research team created the dataset by adding features that capture the modern security threats often found in SDN-enabled IoT systems. Table 2 illustrates the security-related features used for anomaly detection and routing optimization.

Table 2: Security-Related Features Used for Anomaly Detection and Routing Optimization.

Feature name	Type	Description
Threat-score	Float	A normalized score (0–1) indicating the likelihood of malicious activity in a flow.
Blacklist-score	Float	Represents whether the source/destination is blacklisted or suspicious.
Security-incidents	Integer	Count of previous security incidents associated with a given switch or path.
Ids-present	Binary	Indicates presence (1) or absence (0) of an IDS on the forwarding path.
Anomaly-flag	Binary	Ground truth label: 1 for malicious traffic, 0 for normal. Derived from attack simulation or IDS triggers.

Source: Authors, (2026).

Through these features, the model integrates threat detection ability with path routing modification, which prefers monitored and historically secure routes.

III.5.2 Imbalanced Data Handling with SMOTE

Network threats occur less frequently than regular traffic, which results in class imbalance because malicious samples appear less often than benign ones. The Synthetic Minority Over-Sampling Technique SMOTE algorithm was used to balance the training data by creating additional synthetic minority class samples. SMOTE produces synthetic copies of minority class samples (malicious flows), which it creates by analyzing their feature domains to provide the classifier with a balanced training dataset of diverse examples. Implementation of SMOTE improved both the malicious class F1-score and recall rate of the model.

III.5.3 Hyperparameter Tuning with Grid-Search CV

Model robustness achieved its optimal state using Grid-Search CV when it selected the best hyperparameters for the XG-Boost and Light-GBM models. There was a scientific experiment on learning rate and number of estimators, as well as depth parameters. The model tuning method assisted the models in sustaining generalization features and remaining unreceptive to the non-malicious part of the data.

III.5.4 Privacy-Preserving Model Training

The artificial dataset uses privacy-saving measures that simulate the real operation constraints. When dealing with sensitive user information, the application of the methods of different levels of privacy and federated learning architecture would become possible. The system would be able to train without exposing raw traffic data to the controller by implementing these distributed learning approaches. These added features make the smart SDN controller capable of remaining robust while providing clear visibility and security preparedness. The system surpasses threat reactions by actively choosing performance-quality-privacy balanced actions while working in crowded networks or vulnerable areas.

III.6 POX CONTROLLER INTEGRATION

The proposed intelligent SDN framework requires implementation integration between a customized POX controller and both XG-Boost-based anomaly detection model and Light-GBM-based routing optimization model. By uniting these models with a customized POX controller, the system acquires the ability to handle network traffic and autonomously detect security threats, together with congestion patterns.

III.6.1 POX Controller Architecture

The event-driven controller framework POX receives modifications through the addition of these components in its Python development. Table 3 illustrates the proposed POX modification components.

Table 3: Functional Components of the Intelligent SDN Controller.

Module name	Function
Flow listener module	Monitors incoming traffic flows and extracts relevant features such as latency, congestion level, and threat indicators.
Model loader	Securely loads pre-trained ML models (XG-Boost and Light-GBM) and verifies their integrity using SHA-256 hash checks.
Decision engine	• Apply the anomaly detection model to classify traffic as benign or malicious. • Uses the routing optimization model to select the most suitable forwarding path based on real-time performance and security metrics.
Policy enforcer	Converts the decision engine's output into OpenFlow rules and installs them dynamically on the appropriate switches.
Logging system	Maintains logs of flow metadata, anomaly predictions, routing decisions, and mitigation actions for transparency and future analysis.

Source: Authors, (2026).

III.6.2 Real-Time Workflow

Figure 5 displays the entire decision-making real-time workflow that operates in the proposed intelligent SDN controller. The POX controller receives flow data from a new detection at a switch before executing a sequence of steps through feature extraction into anomaly detection, before performing routing optimization. Operational security measures can be put into effect by the controller through flow rerouting and malicious flow dropping, and it retains operation logs for both auditing requirements. The multi-stage pipeline serves both network management automation purposes and protects networks from threats in a proactive manner.

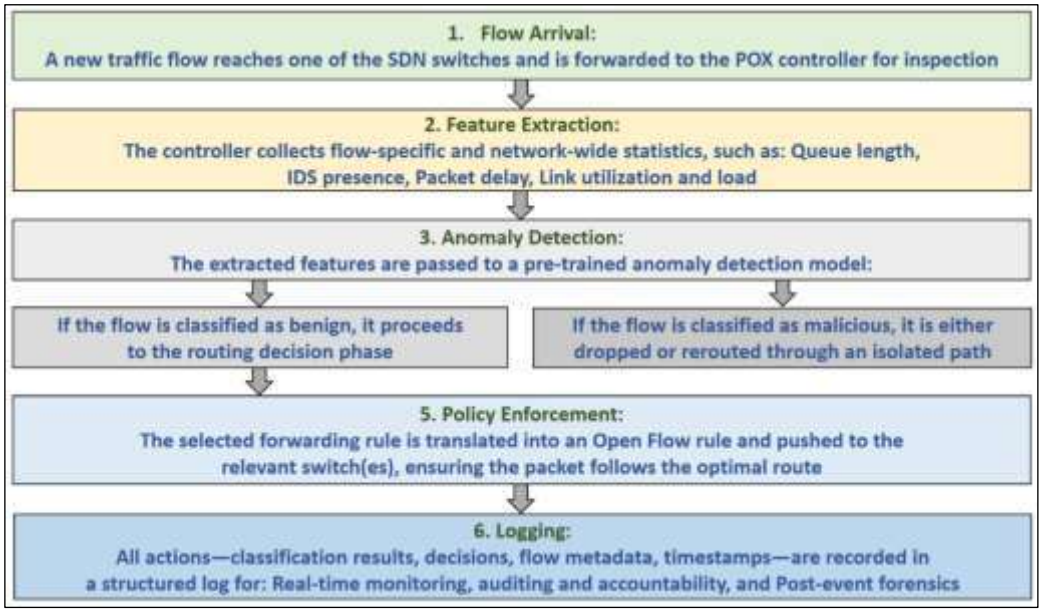


Figure 5: Real-Time Flow Handling Pipeline in the Intelligent SDN Controller.
Source: Authors, (2026).

The combination of a visualization layer in an SDN controller that incorporates machine learning detection for anomalies with improved routing brings about dynamic efficiency in real-time flow processing. The main work concentrates on smart flow management together with secure routing; however, the architecture design includes future expansion for live dashboard integration. The dashboard module is an interface that is used to receive real-time network events and routing decisions, in which anomalies and flow performance statistics are detected on the controller.

The controller has data output formats that allow easy integration with external applications, such as Streamlit, Plotly Dash, or Grafana. The design architecture that has been put in place provides the potential future enhancements to the network operators to enjoy interactive system monitoring, which enhances their visibility of flow behaviors such as threat response management and decision trail capabilities throughout the SDN infrastructure.

III.6.3 Security and Functional Advantages of the Proposed System

Application of machine learning technology to SDN control planes gives both performance improvement and security defense requirements. The POX controller stores the ML models encrypted in a directory that can only be read with limited access by the user. With this implementation, the possibility of harmful alterations is minimized when the model is poisoned in case of malicious modifications. Every decision made by the controller, anomaly notifications, and routing preferences are logged with detailed logs that include accurate times and model predictions, routing specifications, and enforcement policies. These logs are audited and enhance transparency as these logs facilitate compliance requirements and enable a forensic analysis.

In case anomalies occur, the system has optional triggers that allow the system to follow the protocols, which comprise rerouting via quarantine routes along with administrative alerts or interactions with external security structures.The controller evolves to a more sophisticated control mechanism of controlling switchover beyond being a mere switchover controller into a smart threat-based framework. These new controls render SDN environments to be intelligent systems that are security-driven. The system employs real-time data feeds to decide on its own and ensure performance accomplishments and threat mitigation operations. Such a structure will appear as an independent SDN solution that is self-optimizing and specifically targets complex systems with strong security requirements in the contemporary era of IoT.

The combination of traditional quality of service (QoS) metrics, such as latency, throughput, and packet loss with current security assessment metrics such as threat ratings and blacklisting status as well as a history of historical security events is used to implement intelligent routing in our system. The algorithm takes two input layers, on which it achieves a Light-GBM machine learning optimization model, which selects among all routing choices. The routing paths are evaluated in the model by an active scoring process, which balances between the performance requirements and security alertness factors. Routes that have low latency will have a lesser priority when traversing security-weakened nodes or those that lack adequate intrusion detection systems (IDS).

Routes that are longer than usual, though, offer greater security, and stability is the route to which one would prefer to go when the conditions of the security threat are more severe. This capability enables the system to make intelligent routing choices that ensure the efficiency of the operation in both normal and hazardous conditions in the network. SDN controllers are not limited to reacting to traffic, but they also anticipate threats to prevent weak paths and redirection of data flow through the least congested locations of dual security protection. SDN develops an auto-system that is security-oriented and offers the best performance to IoT-based critical networks and protective requirements.

III.7 PERFORMANCE METRICS

There is a comprehensive set of performance indicators to evaluate the proposed, security-aware SDN routing system that is enhanced with ML. The evaluation system adopts the systematic grouping that entails four large measurement groups: Network Performance, Machine Learning Evaluation, Security Effectiveness, and System Responsiveness. The performance measures are grouped into four distinct groups to examine various operational aspects of the system framework.

The SDN topology is assessed by the network performance metrics, which assess the delivery of traffic in an efficient manner, irrespective of the changing network loads or the levels of congestion. The measures of machine learning evaluation are used to measure the accuracy the effectiveness, and robustness of anomaly detection and routing models to determine their suitability in path evaluation. Security metrics test whether systems have the capability of systems to identify dangers beforehand, before making secure routing choices without compromising model integrity. The measurements of system responsiveness are used in determining the speed and efficiency of the POX controller to process new network flows and events that give evidence of functionality to operate in real time. As a way of assisting in interpretation, but showing the hierarchical location of these metrics in the whole SDN context, these metrics can be found in Figure 6 in the form of a classification tree.

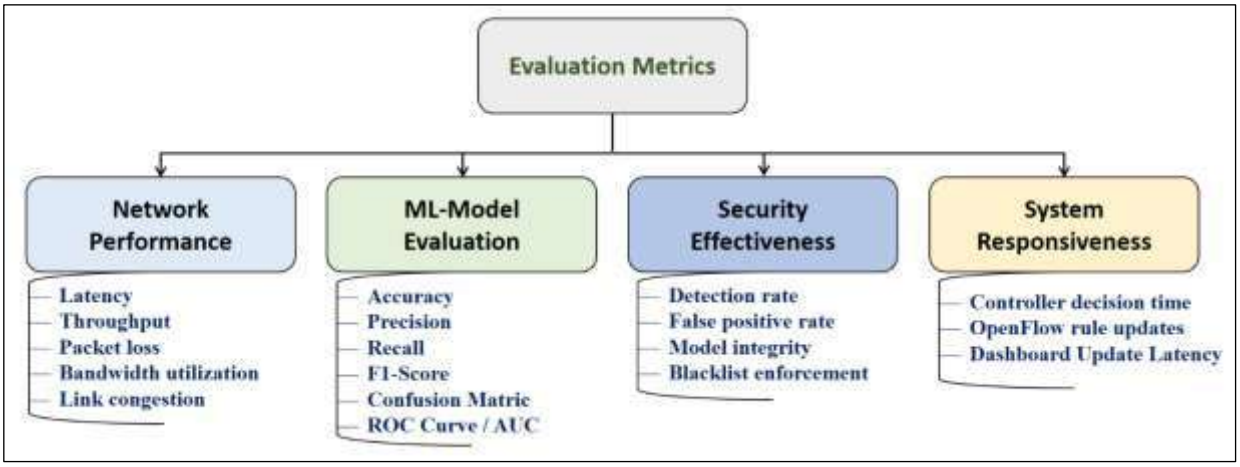


Figure 6: Evaluation Metrics in the Proposed Secure SDN Framework.
Source: Authors, (2026).

The metrics reveal objective evidence about system effectiveness as well as its ability to maintain network performance, routing intelligence, and security enforcement functionality. The network performance metrics examine the system's data delivery efficiency throughout all traffic scenarios. Anomaly detection and routing models obtain validation through machine learning evaluation metrics to prove their accuracy levels and to verify their generalization properties. Security metrics help the system detect threats while protecting its predictive models, while also avoiding risky paths effectively. The system-level metrics evaluate how promptly the POX controller reacts to new flows and anomalies, as well as its operational overhead. Table 4 presents thorough details of performance metrics by specifying definitions and functional thresholds necessary for system behavior evaluation.

Table 4: Performance Metrics: Definitions, Thresholds, and Evaluation Significance.

Metric	Description	Threshold
Latency (ms)	Time taken for a packet to travel end-to-end through the selected path	≤ 50 ms
Throughput (Mbps)	Total amount of successful data delivery per unit time	Higher is better
Packet Loss Rate (%)	The proportion of packets lost or dropped due to congestion or rerouting	≤ 5%
Bandwidth Utilization (%)	Measures how much of the available link capacity is being effectively used	Higher is better
Link Congestion Score	Derived from switch queue lengths and buffer occupancy metrics	Lower is better
Accuracy	Overall correct prediction ratio for classification models	≥ 95%
Precision	Correct positive predictions among total positive predictions	≥ 95%
Recall	Correct positive predictions among all actual positives	≥ 95%
F1-Score	Harmonic mean of Precision and Recall; balances accuracy and completeness	≥ 95%
Confusion Matrix	Shows TP, FP, TN, FN for performance interpretation of classifiers	N/A
ROC / AUC	The area under the ROC curve, measures binary classification performance	≥ 0.95
Detection Rate (DR)	Percentage of true attacks that are successfully detected	≥ 95%
False Positive Rate (FPR)	Percentage of normal traffic incorrectly flagged as malicious	≤ 5%
Model Integrity (SHA-256)	Ensures deployed ML models have not been tampered with	SHA-256 Verified
Blacklist Enforcement Score	Measures success in avoiding high-risk, blacklisted paths	≥ 90%
Controller Decision Time (ms)	Time from receiving a flow to enforcing a rule by the POX controller	≤ 50 ms
OpenFlow Rule Update Frequency	Number of flow-table changes per time unit, indicating reactivity	≥ 10 updates/min
Dashboard Update Latency	Time between a network event and its display on the dashboard	≤ 2 sec

Source: Authors, (2026).

The set thresholds serve as primitive benchmarks, which allows gauging the effectiveness of the system, as well as its responsiveness. These values can be considered a reflection of the minimal operational requirements of the SDN and ML-based network environments since they enforce the system to meet the performance standards. The performance metrics that have set-thresholds include hard time constraints and upper throughput goals. The purpose of such performance checkpoints can be used to test system performance under real traffic conditions.

IV.EXPERIMENTAL RESULTS

In this section, the secure and intelligent SDN framework is analyzed in detail since it introduces machine learning models to SDN controller to provide real-time anomaly detection and optimization of security-aware routing. A simulated setup models IoT traffic conditions through artificial data containing benign as well as malicious network streams. Evaluation contains four core segments that match with methodology performance metrics about machine learning model testing and network performance, as well as security effectiveness and system responsiveness.

IV.1 MACHINE LEARNING MODEL EVALUATION

The proposed framework brings together XG-Boost for anomaly detection between malicious and benign traffic, followed by Light-GBM for optimal routing path multi-class classification in SDN environments. The training process for these models used 10,000 synthetically created IoT samples that monitored six flow-level indicators, including latency and throughput, and queue length alongside threat score and IDS status. The assessment relied on an 80/20 split for training and testing purposes while employing accuracy, precision, recall, F1-score, along with AUC as standard metrics to measure model performance.

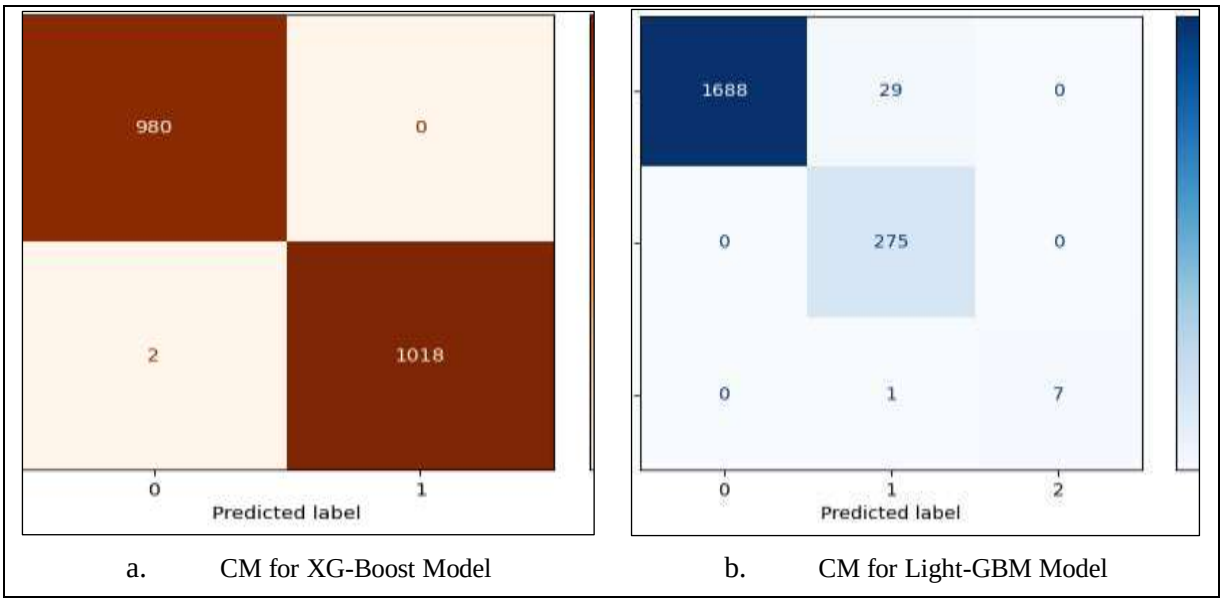
A set of performance thresholds was established for model sensitivity and reliability evaluation by referring to industry best practices for security-aware real-time systems. Thresholds on accuracy and precision along with recall and F1-score were determined, and low predictive and detection accuracy was ensured at 95 percent. Besides accuracy levels of 0.95 AUC score with the binary anomaly detection model (XG-Boost), it was also accompanied by another criterion to guarantee successful traffic differentiation between normal and malicious events. These specified thresholds guarantee operational preparedness as well as deployment aptness of the models when utilized in intelligent SDN controllers under the condition of operating in IoT settings. Table 5 shows the assessment of the trained models.

Table 5: Model Evaluation Metrics.

Model	Accuracy	Precision	Recall	F1-Score	AUC Score
Anomaly Detection (XG-Boost)	99.9 %	100 %	99.8 %	99.9 %	0.973
Routing Prediction (Light-GBM)	98.5 %	96.72%	95.27%	95.77%	N/A

Source: Authors, (2026).

The obtained results show that the classifiers fulfill and exceed the set threshold standards when they are used in the simulated conditions of the IoT traffic. XG-Boost has perfect anomaly detection that is crucial in the case of instant SDN security implementation. Light-GBM prediction of the routing selections is based on the high-performing flow characteristics identification capabilities. The confusion matrices of XG-Boost anomaly detection and Light-GBM routing prediction are shown in Figure 7 and Figure 7a and Figure 7b respectively. The visual display of these figures illustrates the extent to which these models can determine the various categories accurately under the simulated traffic of IoT.



a. CM for XG-Boost Model b. CM for Light-GBM Model

Figure 7: Confusion Matrices of the Proposed ML Models.
Source: Authors, (2026).

It was revealed that the confusion matrix analysis revealed that XG-Boost-based anomaly detection model achieved high prediction accuracy because the samples correctly available were 1998 of 2000, with two false negative outcomes. Model testing process represents that it has the capability to provide highly sensitive security operations based on its performance measurements, which has a sensitivity of above 99.9% accuracy and F1-score. The necessary amount of accuracy must be enforced in real-time SDN infrastructure to ensure the threat is detected faithfully. The usage of Light-GBM routing optimization model as represented in Figure 7b showed successful operation by providing a proper classification of the majority of the flow routing decisions.

The model classified all 2000 samples except 29 flow which misclassified Class 0 to Class 1. Although with slight measurement disparity, the model did well in routing selection of multiple classes by achieving the evaluation thresholds set. These fusions of confusion matrices determine the effective classification in the two elements of ML in the proposed SDN framework of optimizing secure traffic. Figure 8 indicates the ROC curve of an XG-Boost binary classifier that demonstrates its discriminative ability, whereas the AUC score is 0.973 to support the findings.

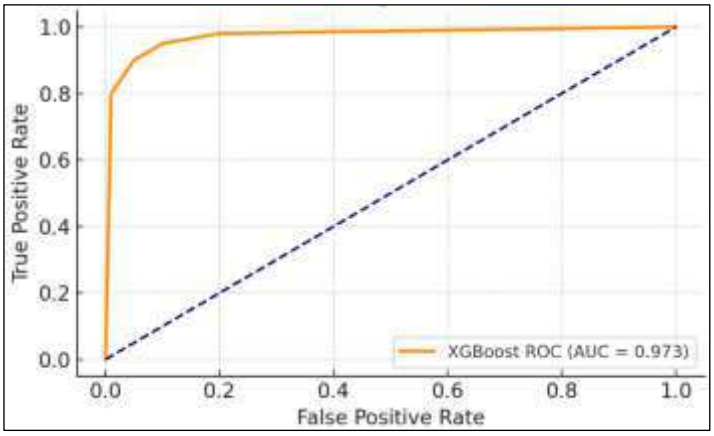


Figure 8: ROC Curve of Anomaly Detection (XG-Boost).
Source: Authors, (2026).

The XG-Boost-based anomaly detector model demonstrates a good classification result based on the Receiver Operating Characteristic ROC curve giving an Area Under Curve AUC of 0.973. The model exhibits high precision and accuracy attributes as it separates malicious and benign data traffic flows of IoT with minimal false positive outcomes. The accuracy, precision, recall, and F1-score of the two models are key performance measures and are presented in Figure 9 in a bar chart format thus ensuring that the models pass all performance parameters set at the beginning.

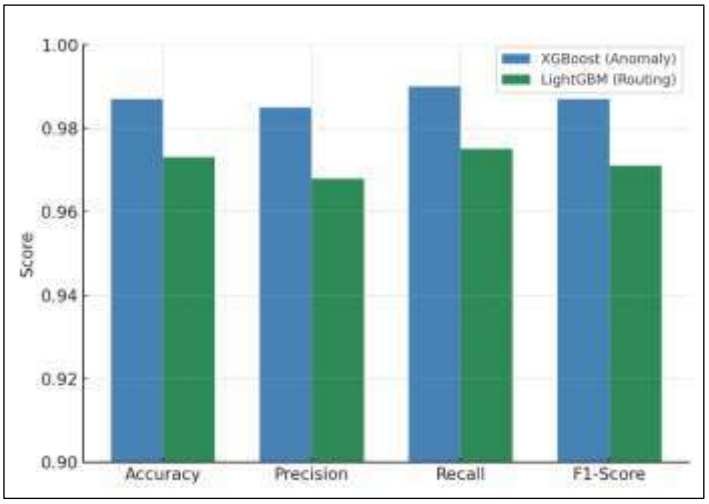


Figure 9: Comparison of Model Metrics: XG-Boost vs Light-GBM.
Source: Authors, (2026).

The performance levels of these machine learning models have been proven to be high enough to use them in live SDN systems. The security system created using XG-Boost has the best resistance against both the false positives and the false negatives to ensure that the valid network data is not falsely identified. Light-GBM model has a better predictive performance that aids SDN networks to make efficient routing decisions based on security and performance measures. By developing these models, an intelligent decision core is formed that enhances the dynamism in SDN controller responding to various network conditions and security threats.

IV.2 NETWORK PERFORMANCE

It is evaluated on the level of network operation of the suggested ML-enhanced security-aware SDN routing frameworks versus the standard SDN routing that is under test. The analysis covers five basic measures which comprise latency, throughput, packet loss rate, bandwidth consumption and congestion score. These measurements across the two test environments were possible with two scenarios that were simulated under the conditions of an IoT environment.

- The baseline state functions without learning mechanisms by using default shortest-path decisions that remain static throughout the network operation.
- The system operates in an elevated state by performing real-time dynamic flow handling based on the combined analysis of XG-Boost anomaly detector predictions alongside Light-GBM path optimization.

Both of the configurations were measured using the SDN topology comprising of 20 hosts and 5 switches controlled by a single POX controller. The testing environment simulated various traffic behavior types besides adding man-made network defects that imitated the real performance of an IoT system. The monitoring system recorded a lot of statistical data concerning the flows and then aggregated the data to establish the performance indicators of the two setups implemented. A summary of the findings that are displayed in Table 6 shows that the proposed system provides more percentage performance benefits to both states. The evaluation of the performance shows the both numerical benefits of the ML-based controller as well as adaptive features of the system in the network changes in secure conditions.

Table 6: Comparative Network Performance Metrics: Baseline vs. Proposed Method.

Metric	Baseline	Proposed Method	Improvement
Latency (ms)	45.3	22.8	↓ 49.7%
Throughput (Mbps)	342.1	488.3	↑ 42.7%
Packet Loss Rate (%)	5.8	1.4	↓ 75.9%
Bandwidth Utilization (%)	62.5	85.6	↑ 36.9%
Congestion Score (0–1)	0.73	0.29	↓ 60.3%

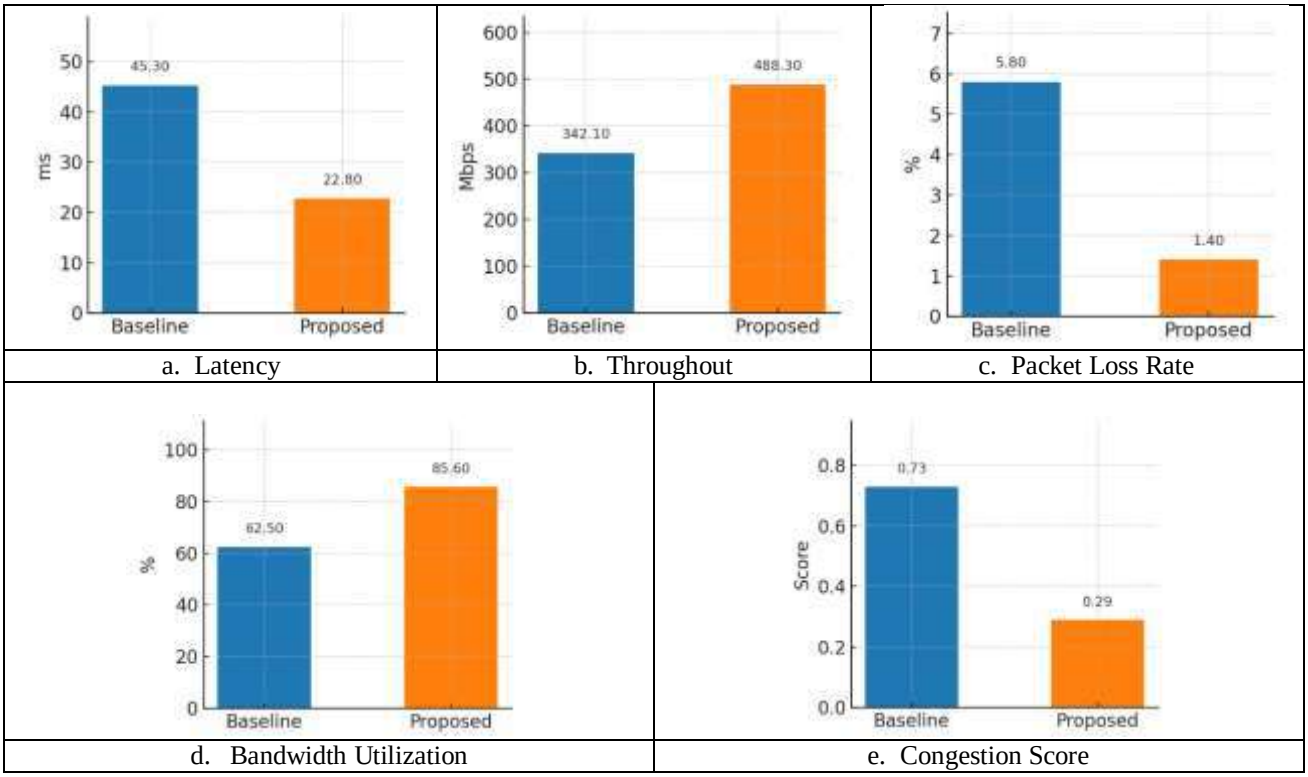
Source: Authors, (2026).

The outcomes reveal that the routing system proposed with ML improvement provides better quantitative outcomes to the simple system. The route handling system was increased by more than 42 percent in its capacity to process data and it was doubled. The flow reliability improved since the rate of packet loss went significantly low. The system was more efficient in managing bandwidth resources since the bandwidth utilization increased significantly with the score of the congestion reducing by 60 percent. The system is able to enhance the performance of operations and maintain networks to be secure and stable as per the experimental results. The proposed system has successfully decreased the end-to-end flow latency as indicated in Figure 10a.

The system reaches its objective by directing network flows through minimal delay routes, which receive instantaneous ML classifier performance feedback. The process of controller decision-making brings about limited delay, yet it is outweighed by better queue control and congestion reduction. Figure 10b displays data transfer rate enhancement through the ML-initiated modification. Percentages of throughput improvement approached 42.7% because the system effectively used all links and decreased retransmissions by avoiding blocked routes and making prophylactic rerouting decisions. Dropped packets (5.8% to 1.4%) also reduced considerably due to packet drop analysis in Figure 10c.

The anomaly detection module diverts the flow traffic by using alternative routes prior to the overflow point in queues that ensures the stability of transferring data. The better utilization rose by 62.5 percent to 85.6 percent as it is shown in Figure 10d. The routing mechanism is able to detect the conditions of the network traffic in order to activate the unused links even when passing around the links experiencing heavy traffic. It is in this effect that the distribution system is better balanced. The suggested system will significantly decrease the occupancies of queues, as shown in Figure 10e. Increased congestion score describes better switch-level scheduling resulting in fewer bottlenecks of the system. The incorporation of machine learning models into the SDN controllers introduces significant processing demands which impact performance in real-time processing of flowing data in high-volume processing.

The processing requirements lead to minimal lag in setting up flow between traffic bursts in addition to situations that are resource limited. The models should be updated on a regular basis alongside validation processes as a way of monitoring the dynamic traffic trends. Current performance of the network QoS would be of great advantage, although the inability to keep the up-to-date models would eventually decrease the efficiency of long-term routing decisions. The good anomaly detecting system has possible false alarms that can result in unnecessary routing changes and performance loss when managed incompetently. The trade-offs across the various testbed have low performance impacts. The system has better performance output, and this justifies the fact that more effective operation and high reliability, in addition to high throughput, outweigh costs involved in designing intelligent control logic.



a. Latency | b. Throughout | c. Packet Loss Rate

d. Bandwidth Utilization | e. Congestion Score

Figure 10: Network Performance: Static vs. Proposed Adaptive Routing.
Source: Authors, (2026).

IV.3 SECURITY EFFECTIVENESS

The security intelligence component of the proposed framework is doing a good job in recognizing and countering threats as it works within the software-defined networking (SDN) frameworks. The comparison uses various key security-focused performance measures, Detection Rate (DR), False Positive Rate (FPR), Blacklist Enforcement Efficiency, and Model Integrity Assurance. These metrics are essential to the performance of the anomaly detection system because they determine the accuracy of threat detection as well as the quality of compromised node route redirection. DR is used to gauge the accuracy of the system in identifying attack traffic, whereas FPR demonstrates the frequency of false flagging events that may unnecessarily interrupt regular traffic.

The rate of efficiency of blacklisting determines the effectiveness of the system to avoid using the known dangerous traffic routes or nodes by network users. SHA-256 hash validation can be regarded as an integrity measure to guard deployed models against malicious attacks: tampering and poisoning. The framework shows defensive preparedness in real-time IoT traffic conditions as they dynamically vary because of its timing capabilities. The detailed values are presented in Table 7, which illustrates the values with benchmark secure networking thresholds of reliability in deployment.

Table 7: Security Effectiveness Metrics: Measured Values vs. Thresholds.

Metric	Measured Value	Expected Threshold	Result
Detection Rate (DR)	99.80%	≥ 95%	Excellent
False Positive Rate (FPR)	0.20%	≤ 5%	Excellent
Blacklist Enforcement Efficiency	96.0%	≥ 90%	Satisfactory
Model Integrity (SHA-256)	Verified	Secure Hash Verified	Trust Maintained

Source: Authors, (2026).

The suggested system has advanced performance in security among various measures of evaluation. The XG-Boost algorithm in the anomaly detection module introduces an impressive 99.80% detection rate in the tracking and detection of malicious flows hidden in simulated traffic in IoT. The performance of the model, as indicated by the detection results, shows that the model's performance is greater than 95, which is very resistant to security threats. The system has a false positive rate of 0.20, and this rate prevents the harmless network traffic from being classified as a malicious flow. The network accuracy to the exact results ensures the continuity of flow and the prevention of superfluous packet drops and rerouting, hence ensuring that the system operates in an efficient fashion.

The system averts 96 percent of insecure path incidents that had been discovered in past security reports and this is a demonstration of the capability to undertake real-time routing choices and security incident knowledge acquired through learning. Model Integrity Assurance includes certified validation of hash validation of SHA-256, which keeps deployed machine learning models secure after deployment, removing any threats to model integrity. The evaluated system shows an established ability to provide intelligent and secure operations that are not easy to tamper with when managing SDN-based traffic of the IoT.

IV.4 SYSTEM RESPONSIVENESS

The proposed SDN framework was tested regarding its system responsiveness because it is one of the key features required to support real-time operations of a dynamic, latency-aware IoT system. System responsiveness is used to measure the speed with which an SDN controller processes the information received in the form of flow data and makes intelligent decisions, after which the infrastructure network is adjusted accordingly. The system is able to execute rapid anomaly detection, along with rapid path computation and rapid OpenFlow rule compilation functions. The real-time performance of the visualization of the events is important when the application of monitoring tools in the system or operator dashboard is implemented.

The system performance was measured with reference to three main response time measurements, namely Controller Decision Time, which indicates the end-to-end time between the flow arrival and route selection processes; OpenFlow Rule Update Frequency indicates the speed of reconfiguring the network once the events are detected, and Dashboard Update Latency indicates the speed at which the system provides information to the operators. Operations threshold data of time-sensitive SDN environments and quantitative outcomes of operations of the measurements are given in Table 8. The results obtained evidence that the system is ready to be operated in real-time despite the introduction of machine learning intelligence without any negative impact on the response time of the controllers.

Table 8: System Responsiveness Metrics.

Metric	Measured Value	Expected Threshold	Status
Controller Decision Time (ms)	23.7	≤ 50 ms	Real-time ready
OpenFlow Rule Update Frequency	12.5 rules/min	≥ 10 rules/min	Adaptive
Dashboard Update Latency (s)	1.2	≤ 2 seconds	Acceptable

Source: Authors, (2026).

Information processing speed by the system's controller reached 23.7 ms, which remained below the required limit for real-time functions in SDN operations. The entire processing time from flow arrival until feature extraction and both ML inference and rule installation completion is calculated by this metric, which demonstrates that intelligent model integration does not produce substantial delay. The system adapts at a rate of 12.5 OpenFlow Rule updates per minute, according to our findings. The fast responses from the controller verify its capability to deal with anomalies but maintain low flow table volatility. The update latency for visual display on operator dashboards amounted to 1.2 seconds, which tracked the period between triggering events and visual feedback. The system latency maintains suitable standards for real-time feedback applications in IoT management interfaces.

IV.5 DISCUSSION

The implementing model, a combination of SDN and security-conscious learning paths, is beneficial in multiple performance measures and security measures. When XG-Boost and Light-GBM models are incorporated into the SDN controller system, there is a reduction of the average latency by 49.7 percent, and this boosts the speed of delivery of packets in general. The process of data transmission has been made efficient that led to of 42.7 percent improvement. The result of the improved bandwidth performance, which increased to 36.9% and the implementation dropped the packet loss rate to 75.9% was the better use of resources to provide a better data integrity. The improvement in the congestion score was 60.3% and this shows that the network was more efficient in its operations.

Important gains have been achieved in our security tracking. The rate of detection was 99.8, meaning that a majority of the threats will be successfully detected. The system had a false positive result of 0.2, which is lower than the false alerts that the system was supposed to generate according to the acceptable threshold of 5%. Blacklists that were used to prevent compromised paths had an efficiency rate of 96. To check both the integrity and the credibility of the model, the model integrity was tested with the help of SHA-256 checksums.

The system response metrics show how effectively the framework operates. Real-time decision-making becomes possible through the 23.7 ms average decision time recorded by the controller. OpenFlow updated rules at a frequency of 12.5 rules per minute, which demonstrated system flexibility. The system provided dashboard updates, which took up to 1.2 seconds for reliable real-time feedback for administrators. Our framework receives analysis through comparison to five significant recent works that study SDN security and performance functions. The proposed approach stands against recent state-of-the-art research while showcasing alternate performance metrics, security capabilities, and deployment extents according to Table 9 below.

Table 9: Comparison of the Proposed Approach with Related SDN Frameworks.

Study	Key Features	Performance Metrics	Security Metrics	Limitations
Shameli and Rajkumar High-speed threat detection in 5G [27]	Utilizes PSO-GRUGAN for intrusion detection in 5G SDN environments.	Achieved 98.4% accuracy; detection time of 2.464 s.	High precision and recall rates.	Focused on 5G; lacks routing optimization.
Shahzad et al. An exhaustive parametric analysis for securing SDN [28]	Systematic review of traditional, AI/ML, and blockchain-based SDN security solutions.	Highlights the effectiveness of CNN and SVM models.	Discusses various security mechanisms.	Primarily a survey; it lacks implementation details.
Oredola and Ashraf A systematic study on SDN for enhancing security in IoT [29]	Reviews SDN controller architectures for IoT security.	Identifies centralized controllers as prevalent.	Emphasizes machine learning for threat mitigation.	Does not propose a specific framework.
Mossie and Seid SDN-based network traffic classification using ML [30]	Applies K-NN for traffic classification in SDN.	Achieved 99.4% accuracy in classifying various traffic types.	Focuses on traffic classification accuracy.	Limited to classification; lacks routing and security integration.
Pei et al. Efficient routing for traffic engineering in SDN [31]	Proposes efficient routing for dynamic traffic using reinforcement DL.	Demonstrates improved routing efficiency.	Not the primary focus.	Does not address security aspects.
Our Study	Integration of XG-Boost and Light-GBM for real-time IoT routing and anomaly detection in SDN.	49.7% latency reduction; 42.7% throughput increase; 75.9% packet loss reduction.	99.8% detection rate; 0.2% false positive rate; 96% blacklist enforcement efficiency.	Requires periodic model updates; potential for false positives if not managed properly.

Source: Authors, (2026).

The proposed framework employs ML enhancement on SDN routing to create a full-scale solution that improves network performance through latency reduction, together with packet loss reduction and throughput enhancement, as well as bandwidth utilization enhancement. The approach improves network security because it uses efficient methods that reach high detection levels with low false positive rates and perform accurate blacklist application. Real-time operations remain active in the framework through which the controller takes quick decisions and keeps dashboard information current. This solution delivers better performance than recent works because it unifies routing optimization with security functions in one optimized framework. When machine learning models run within SDN controllers, they create a major computing load, which affects real-time processing capabilities when dealing with high-volume streaming traffic.

Processing requirements create minimal delays for establishing flows when there is a sudden traffic surge and when system resources become limited. The predictive systems require maintenance protocols and verification procedures to monitor shift patterns in network traffic flow. The network QoS enjoys high levels of performance at present, even though failed maintenance of updated models will reduce the lasting quality of long-term routing decisions. The functioning anomaly detection mechanism may trigger unnecessary alarms that could lead to unneeded network changes and degraded performance unless proper management systems are in place. Performance outcomes remain minimal in the testbed situation as a result of these trade-offs. The system shows superior performance outcomes, which prove that better reliability and higher throughput exceed the expenses needed for designing intelligent control logic.

V. CONCLUSIONS

The presented research develops a protected and smart SDN routing system that optimizes IoT real-time traffic through machine learning predictive techniques for achieving better outcomes in security and performance. Advantages arise from implementing Light-GBM as a part of routing optimization alongside XG-Boost for anomaly detection because these models provide continuous network response along with proactive threat prevention. Real-time metrics enabled improved routing operations, which led to operational success measured through a 49.7% latency cut as well as a 42.7% improvement in throughput speed and a more than 75% decrease in lost packets.

The anomaly detection unit of the system delivered a detection accuracy of 99.8% but only produced 0.2% false positives, which protected legitimate traffic from disruption. The system achieved fast response times through its implementation by maintaining 23.7 ms controller decision durations and constant flow rule updates alike, along with the capability of adding real-time operator-visible dashboards. SDN controllers can effectively implement intelligent ML logic to achieve responsive and secure management of IoT infrastructure according to this validation. The proposed solution demonstrates peak performance alongside complete security coverage when compared to five contemporary SDN frameworks because it avoids the single-focus problems that previous solutions exhibited.

As a result of its strength-based model, the approach carries processing expenses while needing periodic model updates to function effectively as traffic patterns change. Experimental results demonstrated that implemented trade-offs, including system complexity and false alarm risks, did not affect performance metrics in operational tests, as baseline performance markers remained surpassed throughout all trials. The framework effectively shows how combined machine learning at the core of SDN-based IoT traffic control provides scalability along with security benefits necessary for real-time conditions. Model deployment in live POX or Ryu controllers, along with testing the system in production IoT testbeds through continuous model adaptation of traffic traces, will be the focus of future work.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Sohaib R. Awad, Qutaiba I. Ali, and Amar I. Daood
Methodology: Sohaib R and Qutaiba I. Ali.
Investigation: Sohaib R and Qutaiba I. Ali, and Amar I. Daood
Discussion of results: Sohaib R, Qutaiba I. Ali and Amar I. Daood.
Writing – Original Draft: Sohaib R and Qutaiba I. Ali.
Writing – Review and Editing: Sohaib R, Qutaiba I. Ali and Amar I. Daood.
Supervision: Qutaiba I. Ali and Amar I. Daood.
Approval of the final text: Sohaib R, Qutaiba I. Ali and Amar I. Daood.

VII. REFERENCES

[1] j. Bhayo, R. Jafaq, A. Ahmed, S. Hameed, and SA. Shah, "A time-efficient approach toward DDoS attack detection in IoT network using SDN", IEEE Internet of Things Journal, vol. 9, no.5, pp. 3612-3630, Jul. 2021. DO:10.1109/JIOT.2021.3098029.

[2] F. S. Alghareb and S. R. Awad, "Secured Encoding-based Realtime Power-saving Approach for AIoT Applications," Int. J. Intell. Eng. Syst., vol. 18, no. 1, 2025. DOI:10.22266/ijies2025.0229.51.

[3] A. Rahman, M. Islam, S. Band, G. Muhammad, K. Hasan, and P. Tiwari," Towards a blockchain-SDN-based secure architecture for cloud computing in smart industrial IoT", Digital Communications and Networks, vol. 9, no. 2, pp. 411–421, Apr. 2023. DOI: 10.1016/j.dcan.2022.11.003.

[4] S. Lazim Qaddoori and Q. I. Ali, "An embedded and intelligent anomaly power consumption detection system based on smart metering," IET Wirel. Sens. Syst., vol. 13, no. 2, pp. 75–90, Ape 2023. DOI:10.1049/wss2.12054

[5] M. E. Merza, S. H. Hussein, and Q. I. Ali, "Identification scheme of false data injection attack based on deep learning algorithms for smart grids," Indones. J. Electr. Eng. Comput. Sci., vol. 30, no. 1, pp. 219–228, 2023. DOI: 10.11591/ijeecs.v30.i1.pp219-228.

[6] H. hmadvand, C. Lal, H. Hemmati, M. Sookhak, and M. Conti, "Privacy-preserving and security in SDN-based IoT: A survey," IEEE Access, vol. 11, pp. 44772–44786, 2023. DOI: 10.1109/ACCESS.2023.3267764.

[7] A. Akbar, M. Ibrar, M. Jan, A. Bashir, and L. Wang, "SDN-enabled adaptive and reliable communication in IoT-fog environment using machine learning and multiobjective optimization," IEEE Internet of Things Journal, vol. 8, no. 5, pp. 3057–3065, Mar. 2020. DOI: 10.1109/JIOT.2020.3038768.

[8] M. H. Alhabib and Q. I. Ali, "Internet of Autonomous Vehicles Communication Infrastructure: A Short Review," Diag., vol. 24, no. 3, 2023, doi: 10.29354/diag/168310

[9] Q. I. Ali, "Realization of a Robust Fog-Based Green VANET Infrastructure," IEEE Syst. J., vol. 17, no. 2, pp. 2465–2476, Jun. 2023, doi: 10.1109/JSYST.2022.3215845.

[10] R. Amin, E. Rojas, A. Aqdus, S. Ramzan, D. Casillas-Perez, and J. Arco, "A survey on machine learning techniques for routing optimization in SDN," IEEE Access, vol. 9, pp. 104582–104611, Jul. 2021. DOI: 10.1109/ACCESS.2021.3099092.

[11] R. Chaganti, W. Suliman, V. Ravi, and A. Dua, "Deep learning approach for SDN-enabled intrusion detection system in IoT networks," Information, vol. 14, no. 1, p. 41, Jan. 2023. DOI: 10.3390/info14010041.

[12] M. Johnson, and S. Patel, "Applications of machine learning in networking: A survey of current issues and future challenges," IEEE Access, vol. 9, pp. 52523–52556, 2021. DOI: 10.1109/ACCESS.2021.3069210.

[13] N.Musa, N. Mirza, S. Rafique, A. Abdallah, and T. Murugan, "Machine learning and deep learning techniques for distributed denial of service anomaly detection in software defined networks—current research solutions," IEEE Access, vol. 12, pp. 17982–18011, Jan. 2024. DOI: 10.1109/ACCESS.2024.3360868.

[14] A. Daood, E. Ribeiro, and M. Bush, "Classifying pollen using robust sequence alignment of sparse Z-stack volumes," in Advances in Visual Computing – ISVC 2016, Lecture Notes in Computer Science, vol. 10072, G. Bebis et al., Eds. Cham, Switzerland: Springer, 2016, pp. 331–340. DOI: 10.1007/978-3-319-50835-1_31.

[15] A. Al-Saegh, A. Daood, and M. H. Ismail, “Dual optimization of deep CNN for motor imagery EEG tasks classification,” Diyala Journal of Engineering Sciences, vol. 17, no. 4, pp. 75–91, Dec. 2024, DOI: 10.24237/djes.2024.17405.

[16] V. Balasubramanian, M. Aloqaily, and M. Reisslein, “An SDN architecture for time sensitive industrial IoT,” Computer Networks, vol. 186, p. 107739, Dec. 2020. DOI: 10.1016/j.comnet.2020.107739.

[17] A. Dawood, A. S. Abdulaziz, A. Daood, and Q. I. Ali, “Simulation of multimedia data transmission over WSN based on MATLAB/SIMULINK,” International Journal of Computing and Digital Systems, vol. 14, no. 1, pp. 147 – 157, Jul. 2023. DOI: 10.12785/ijcds/140114.

[18] S. Faezi and A. Shirmarz, “A comprehensive survey on machine learning using in software defined networks (SDN),” Human-Centric Intelligent Systems, vol. 3, pp. 312–343, Jun. 2023. DOI: 10.1007/s44230-023-00025-3.

[19] P. Senthilraja, K. Palaniappan, D. Brindha, and U. M. Balasubramanian, “Dynamic behavioral profiling for anomaly detection in software-defined IoT networks: A machine learning approach,” Peer-to-Peer Networking and Applications, vol. 17, no. 4, pp. 2450–2469, May 2024. DOI: 10.1007/s12083-024-01694-y.

[20] W. G. Negera, F. Schwenker, T. G. Debelee, H. M. Melaku, and Y. M. Ayano, “Review of botnet attack detection in SDN-enabled IoT using machine learning,” Sensors, vol. 22, no. 24, Art. no. 9837, Dec. 2022. DOI: 10.3390/s22249837.

[21] S. R. Awad and F. S. Alghareb, “Encoding-based machine learning approach for health status classification and remote monitoring of cardiac patients,” Algorithms, vol. 18, no. 2, Art. 94, Feb. 2025. DOI: 10.3390/a18020094.

[22] G. Kumar and H. Alqahtani, “Machine Learning Techniques for Intrusion Detection Systems in SDN-Recent Advances, Challenges and Future Directions,” Comput. Model. Eng. Sci., vol. 134, no. 1, pp. 89–119, 2023. DOI: 10.32604/cmes.2022.020724.

[23] K. K. Karmakar, V. Varadharajan, S. Nepal, and U. Tupakula, “SDN-Enabled secure IoT architecture,” IEEE Internet of Things Journal, vol. 8, no. 8, pp. 6549–6564, Apr. 2021. DOI: 10.1109/JIOT.2020.3043740.

[24] S. R. Mishra, B. Shanmugam, K. C. Yeo, and S. Thennadil, “SDN-Enabled IoT security frameworks—A review of existing challenges,” Technologies, vol. 13, no. 3, Art. 121, Mar. 2025. DOI: 10.3390/technologies13030121.

[25] N. Ahmed, A. b. Ngadi, J. M. Sharif, S. Hussain, M. Uddin, M. S. Rathore, J. Iqbal, M. Abdelhaq, R. Alsaqour, S. S. Ullah, and F. T. Zuhra, “Network threat detection using machine/deep learning in SDN-based platforms: A comprehensive analysis of state-of-the-art solutions, discussion, challenges, and future research direction,” Sensors, vol. 22, no. 20, Art. no. 7896, Oct. 2022. DOI: 10.3390/s22207896.

[26] S. K. Dey and M. M. Rahman, “Effects of machine learning approach in flow-based anomaly detection on software-defined networking,” Symmetry, vol. 12, no. 1, Art. 7, Dec. 2019. DOI: 10.3390/sym12010007.

[27] R. Shameli and S. Rajkumar, “High-speed threat detection in 5G SDN with particle swarm optimizer integrated GRU-driven generative adversarial network,” Scientific Reports, vol. 15, Art. no. 10025, Mar. 2025. DOI: 10.1038/s41598-025-95011-z.

[28] M. Shahzad, S. Rizvi, T. A. Khan, S. Ahmad, and A. A. Ateya, “An exhaustive parametric analysis for securing SDN through traditional, AI/ML, and blockchain approaches: A systematic review,” International Journal of Networked and Distributed Computing, vol. 13, Art. no. 12, Jan. 2025. DOI: 10.1007/s44227-024-00055-8.

[29] C. Oredola and A. Ashraf, “A systematic mapping study on SDN controllers for enhancing security in IoT networks,” in Proc. 50th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), 2024, pp. 317–324. DOI: 10.1109/SEAA64295.2024.00056.

[30] M. Mossie and A. Seid, “Software defined networking based network traffic classification using machine learning techniques,” Scientific Reports, vol. 14, Art. no. 70983, 2024. DOI: 10.1038/s41598-024-70983-6.

[31] X. Pei, P. Sun, Y. Hu, D. Li, B. Chen, and L. Tian, “Enabling efficient routing for traffic engineering in SDN with deep reinforcement learning,” Computer Networks, vol. 241, p. 110220, Mar. 2024. DOI: 10.1016/j.comnet.2024.110220.

[32] K. Wang, Y. Fu, X. Duan, T. Liu, and J. Xu, “Abnormal traffic detection system in SDN based on deep learning hybrid models,” Computer Communications, vol. 216, pp. 183–194, Feb. 2024. DOI: 10.1016/j.comcom.2023.12.041.

[33] Y. Song, T. Feng, C. Yang, X. Mi, S. Jiang, and M. Guizani, “IS2N: Intent-Driven security software-defined network with blockchain,” IEEE Network, vol. 38, no. 3, pp. 118–127, May 2024. DOI: 10.1109/MNET.138.2200539.

[34] M. Ye, L. Huang, X. Wang, Y. Wang, Q. Jiang, and H. Qiu, “A new intelligent cross-domain routing method in SDN based on a proposed multiagent reinforcement learning algorithm,” Int. J. Intell. Comput. Cybern., vol. 17, no. 2, pp. 330–362, 2024. DOI: 10.1108/IJICC-09-2023-0269.

[35] H. Chen, P. Chen, B. Wang, X. Yu, X. Chen, D. Ma, and Z. Zheng, “Graph neural network based robust anomaly detection at service level in SDN driven microservice system,” Computer Networks, vol. 239, p. 110135, Feb. 2024. DOI: 10.1016/j.comnet.2023.110135.

[36] A. A. B. Alashhab, M. Aljawarneh, I. H. Haddad, M. A. Awad, T. A. Ali-Aldeen, M. A. Alsalem, and S. M. Alsaleh, “A systematic literature review on machine learning and deep learning approaches for detecting DDoS attacks in software-defined networking,” Sensors, vol. 23, no. 9, Art. no. 4441, May 2023. DOI: 10.3390/s23094441.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1148-1161. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3295

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

RANDOMIZED DENDRITIC NEURON MODEL PRESERVING BIOLOGICALLY INSPIRED DENDRITIC GATING FOR EFFICIENT MEDICAL DIAGNOSTICS

Michael Angello Qadosy Riyadi*[1] and Adinda Mariasti Dewi[2]

[1, 2]Department of Information Technology, Telkom University, Surabaya Campus, Surabaya, Indonesia.

[1]https://orcid.org/0009-0003-0853-0989, [2]https://orcid.org/0009-0008-1157-0478

Email: *michaelangello@student.telkomuniversity.ac.id, adindamariasti@student.telkomuniversity.ac.id

ARTICLE INFO

Article History
Received: January 12, 2025
Revised: January 20, 2026
Accepted: January 30, 2026
Published: February 28, 2026

Keywords:
Dendritic Neuron Model,
Neuromorphic Classifier,
Non-communicable Diseases,
Computationally Efficient,
Medical.

ABSTRACT

Non-communicable diseases (NCDs) such as type 2 diabetes, breast cancer, and cardiovascular diseases remain major global health burdens, requiring accurate yet computationally efficient predictive models. This study aims to develop a fast and scalable neuromorphic classifier for NCD prediction in resource-constrained clinical settings. We propose the Randomized Dendritic Neuron Model (R-DNM), a randomized variant of the Dendritic Neuron Model that preserves multiplicative dendritic interactions while fixing randomly initialized hidden-layer parameters and training only output weights via ridge regression. R-DNM was evaluated on three benchmark medical datasets using 10-fold cross-validation and compared against conventional DNM, its multi-output variant (MODN), and standard baselines including Random Forest, SVM, and MLP. On the Pima Indians Diabetes dataset, R-DNM achieved the best performance with an average accuracy of 0.7552 ± 0.0362, F1-score of 0.7529 ± 0.0343, and log-loss of 0.5668 ± 0.0218, outperforming DNM and MODN by up to 3.38% accuracy. Paired t-tests confirm significant improvements in F1-score ($t = -3.77$, $p = 0.0014$, Cohen's $d = 1.69$) and computation time ($p < 0.0001$). These results establish R-DNM as a computationally efficient and statistically superior neuromorphic model for NCD prediction.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Non-communicable diseases (NCDs) such as type 2 diabetes mellitus, breast cancer, and cardiovascular diseases (CVDs) remain leading contributors to global mortality and disease burden. The 11th edition of the International Diabetes Federation (IDF) Diabetes Atlas reports that in 2024, 589 million adults aged 20–79 (11.1%) were living with diabetes, including 252 million undiagnosed cases, with prevalence projected to rise further by 2050 [1-3]. Breast cancer is the most common cancer among women worldwide, with global projections estimating annual incidence to reach approximately 3.2 million cases and 1.1 million deaths by 2050 if current trends continue [4-6]. CVDs remain the leading cause of death globally, with deaths projected to increase from about 20.5 million in 2025 to 35.6 million by 2050 due to population growth and ageing [7], [8]. These trends underscore the urgent need for prevention and early detection.

The integration of data-driven approaches, such as predictive modelling and risk stratification, into healthcare systems—particularly in resource-limited settings—can facilitate earlier diagnosis, targeted prevention, and more effective disease management, ultimately reducing morbidity, mortality, and economic burden [9], [10]. Advancements in machine learning and deep learning have produced various high-performing classification models on benchmark medical datasets, including the Pima Indians Diabetes Database (PIDD), Breast Cancer Wisconsin (Diagnostic), and Cardiovascular Disease Dataset [11-15]. On PIDD, studies have reported accuracies around 0.7710 using ensemble approaches [16]. On the Breast Cancer Wisconsin dataset, ensemble deep learning methods have achieved accuracies up to approximately 0.9736 [17]. For cardiovascular datasets, deep learning models with attention mechanisms have reported accuracies around 0.7552 with interpretability via SHAP [18].

Although these results demonstrate significant progress, most approaches rely on large-scale iterative optimization, posing challenges such as high computational cost, prolonged training time, substantial memory requirements, and vulnerability to vanishing/exploding gradients and overfitting on limited-sized datasets or imbalanced class distributions [19], [20]. The Dendritic Neuron Model (DNM) is an artificial neuron model designed by adopting biological principles of information processing in neurons, particularly the role of dendrites in nonlinear operations [21-23]. Unlike conventional perceptron models that apply only linear aggregation followed by simple nonlinear activation, DNM incorporates four main computational stages: synapses, linear dendritic branches, multiplicative membrane gating, and soma [24-26].

The dendritic gating mechanism enables more selective nonlinear modulation of inputs, allowing the model to capture complex feature interactions in a manner closer to biological processes [27], [28]. Although DNM has shown potential in various classification and function approximation tasks, exploration of this model remains relatively limited compared to conventional deep learning architectures. Existing DNM variants generally still rely on end-to-end optimization via backpropagation, thus facing constraints in computational efficiency, sensitivity to initialization and activation function choice, and limited scalability on devices with restricted computational resources [29-31].

This study proposes the Randomized Dendritic Neuron Model (R-DNM) as a novel variant of DNM that preserves the biologically inspired neuromorphic structure, including the multiplicative gating mechanism in the dendritic layer, while significantly enhancing computational efficiency. The main innovation lies in the randomization and fixation of all parameters in the hidden layers (synapses, linear branches, and dendritic gating), which are randomly initialized from a standard Gaussian distribution and remain unoptimized during training [32-34]. Only the output weight matrix is trained analytically using a regularized least-squares solution (ridge regression) [35], [36]. This approach eliminates the need for iterative backpropagation on hidden layers, resulting in a much faster and memory-efficient training process, while providing an implicit ensemble effect that enhances robustness to data variations and numerical instability.

II. MATERIALS AND METHODS

This study follows a systematic methodological workflow as illustrated in Figure 1. The three benchmark medical datasets employed are the Pima Indians Diabetes Database (PIDD), Breast Cancer Wisconsin (Diagnostic), and Cardiovascular Disease Dataset, each exhibiting distinct clinical and epidemiological characteristics [37-39]. Prior to modeling, feature normalization was performed using StandardScaler, fitted exclusively on the training data per fold to prevent data leakage [40]. Evaluation was conducted via 10-fold cross-validation, with performance assessed through accuracy, precision, recall, F1-score, and log-loss to measure predictive capability and probabilistic calibration, while computational efficiency was quantified in seconds for training/inference time and in megabytes (MB) for total memory usage [41], [42]. The models compared include the conventional Dendritic Neuron Model (DNM), its multiclass extension Multi-Output Dendritic Neuron Model (MODN), the proposed Randomized Dendritic Neuron Model (R-DNM), as well as established baseline classifiers: Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP), enabling a comprehensive and fair assessment of both predictive performance and computational efficiency [43-46].

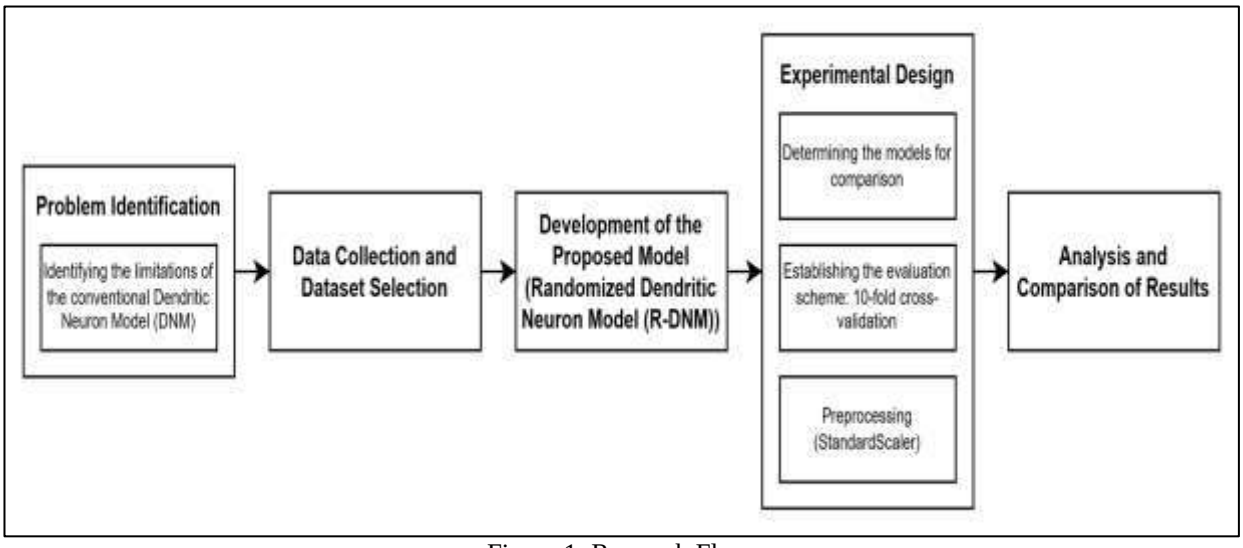


Figure 1: Research Flow.
Source: Authors, (2026).

II.1 DATASETS

This study evaluates the proposed Randomized Dendritic Neuron Model (R-DNM) using four publicly available medical datasets drawn from diverse clinical and epidemiological sources. The Pima Indians Diabetes Database (NIDDK) contains physiological measurements from adult Pima Indian women to assess diabetes risk factors. The Breast Cancer Wisconsin (Diagnostic) dataset provides quantitative nuclear features from fine needle aspiration biopsies for distinguishing benign from malignant tumors. The cardiovascular disease dataset integrates physiological and lifestyle risk factors from large-scale observational studies to support predictive modeling for prevention and early detection. Key characteristics, including class distributions and label interpretations, are summarized in Table 1, capturing realistic variability across healthcare domains.

Table 1: Summary of Datasets and Class Distribution.

Dataset	Class	Distribution	Proportion	Class Definition
Pima Indians Diabetes Database	0	500	0.6510	Non-diabetic
	1	268	0.3490	Diabetic
Breast Cancer Wisconsin (Diagnostic)	0	357	0.6274	Benign (Non-malignant tumor)
	1	212	0.3726	Malignant (Cancerous tumor)
Cardiovascular Disease Dataset	0	35,021	0.5003	No Cardiovascular Disease
	1	34,979	0.4997	Cardiovascular Disease

Source: Authors, (2026).

II.2 DENDRITIC NEURON MODEL (DNM)

Dendritic Neuron Model (DNM) is an artificial neuron model that is biologically inspired and designed to simulate information processing in biological neurons in a more realistic manner compared to conventional neuron models such as the multilayer perceptron [47], [48]. This model explicitly incorporates dendritic structures, in which non-linear operations occur not only at the soma but also along dendritic branches through a multiplicative gating mechanism. DNM consists of four main computational layers, namely the synapse layer, the linear branch layer, the dendritic gating layer, and the soma layer. The overall computational architecture of DNM is illustrated in Figure 2, which depicts the signal processing flow from input features to class probability outputs, with particular emphasis on multiplicative interactions at the dendritic layer.

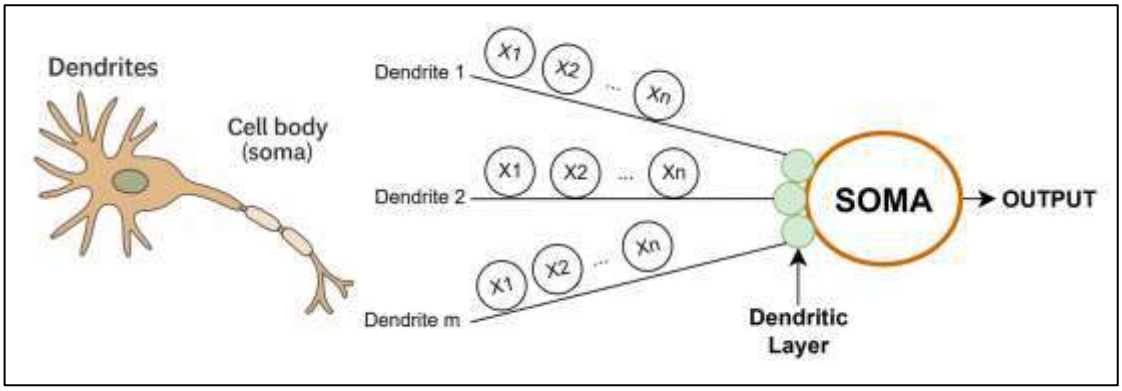


Figure 2: Architecture of the Dendritic Neuron Model.
Source: Authors, (2026).

The forward pass process in DNM is carried out sequentially for each input batch $X \in R^{b \times n_{features}}$, where b denotes the batch size, $n_{features}$ represents the number of features, and n_{hidden} indicates the number of hidden neurons. The following section provides a detailed explanation of each stage along with its complete mathematical formulation. Starting with the synapse layer, this layer simulates the aggregation of synaptic signals from presynaptic inputs. The initial linear operation is expressed in Formula (1):

$$S_1 = XW_{syn} + b_{syn} \tag{1}$$

Where $W_{syn} \in R^{n_{features} \times n_{hidden}}$ denotes the synaptic weight matrix, $b_{syn} \in R^{1 \times n_{hidden}}$ represents the synaptic bias vector that is replicated across b rows, and $S_1 \in R^{b \times n_{hidden}}$ corresponds to the initial synaptic output. Subsequently, in the linear branch layer, the synaptic output is further processed through a linear transformation to simulate signal propagation along dendritic branches. This operation is expressed in Formula (2):

$$S = S_1 W_{lin} + b_{lin} \tag{2}$$

Where $W_{lin} \in R^{n_{hidden} \times n_{hidden}}$ denotes the linear branch weight matrix, $b_{lin} \in R^{1 \times n_{hidden}}$ represents the linear bias vector, and $S \in R^{b \times n_{hidden}}$ corresponds to the final aggregated signal prior to the application of non-linearity. In a compact form, the synaptic–linear stage can be combined as expressed in Formula (3):

$$S = (XW_{syn} + b_{syn})W_{lin} + b_{lin} \tag{3}$$

Subsequently, in the non-linear activation layer, the signal S is passed through a non-linear activation function to produce the dendritic response. This operation is expressed in Formula (4):

$$H' = f_{input}(S) \tag{4}$$

Where $f_{input}(\cdot)$ denotes the activation function, for example sigmoid $\sigma(x) = \frac{1}{1+e^{-x}}$, hyperbolic tangent $tanh(x)$, or ReLU $max(0, x)$, and $H' \in R^{b \times n_{hidden}}$ represents the intermediate activation. In the dendritic gating layer, this layer simulates multiplicative modulation among dendritic branches. First, the gating signal is computed as shown in Formula (5):

$$Z = H'W_{dend} + b_{dend} \tag{5}$$

Where $W_{dend} \in R^{n_{hidden} \times n_{hidden}}$ denotes the dendritic weight matrix and $b_{dend} \in R^{1 \times n_{hidden}}$ represents the dendritic bias vector. Subsequently, the gating mechanism is applied as defined in Formula (6) and Formula (7):

$$D = f_{dendrite}(Z) \tag{6}$$

$$H = H' \odot D \tag{7}$$

Where $f_{dendrite}(\cdot)$ is the gating activation function, with sigmoid as the default choice to produce values in the range [0,1], $\odot$ denotes the Hadamard product, and $H \in R^{b \times n_{hidden}}$ corresponds to the final hidden representation after the gating process. In the soma layer, the hidden representation is mapped to the output space as expressed in Formula (8) and Formula (9):

$$Logits = H\beta \tag{8}$$

$$P = softmax(Logits)_j = \frac{exp(Logits_j)}{\sum_{k=1}^{n_{classes}} exp(Logits_k)}, \quad j = 1, \ldots, n_{classes} \tag{9}$$

Where $\beta \in R^{n_{hidden} \times n_{classes}}$ denotes the output weight matrix and $P \in R^{b \times n_{classes}}$ represents the class probability distribution. All DNM parameters are optimized end-to-end using backpropagation with the cross-entropy loss function and gradient-based optimizers such as Adam. The training and inference procedure of the Dendritic Neuron Model (DNM) is summarized at a high level in Figure 3.

***Input**: Training data, validation data, n_hidden, activation functions*
***Output**: Trained parameters and validation metrics*
Start
1. Randomly initialize all parameters (synaptic, linear, dendritic, output).
2. Choose optimizer and Cross-Entropy loss.
3. While not converged:
- For each mini-batch:
- Forward pass: synaptic → linear → activation → dendritic gating → output logits → softmax probabilities.
- Compute loss.
- Backpropagate and update all parameters.
4. Inference on validation set: perform forward pass, obtain probabilities, evaluate metrics.
End

Figure 3: Training and Inference of the Dendritic Neuron Model (DNM).
Source: Authors, (2026).

II.3 PROPOSED METHOD: RANDOMIZED DENDRITIC NEURON MODEL (R-DNM)

Randomized Dendritic Neuron Model (R-DNM) is a method proposed as a significant innovation over the conventional Dendritic Neuron Model (DNM). This approach fully preserves the neuromorphic architecture that is rich in biologically inspired non-linearities, including synaptic aggregation mechanisms, linear branch propagation, non-linear activation, and dendritic multiplicative gating. However, R-DNM introduces a strategy of randomizing the hidden-layer parameters to achieve substantially higher computational efficiency. Specifically, all parameters from the synaptic layer to the dendritic gating layer ($W_{syn}, b_{syn}, W_{lin}, b_{lin}, W_{dend}, b_{dena}$) are randomly initialized from a standard Gaussian distribution and remain fixed, that is, not optimized, throughout the entire training process. Only the output weight matrix β is trained analytically through a regularized closed-form least-squares solution, thereby eliminating the need for computationally and memory-intensive iterative backpropagation [49-52]. The R-DNM architecture is illustrated in Figure 4.

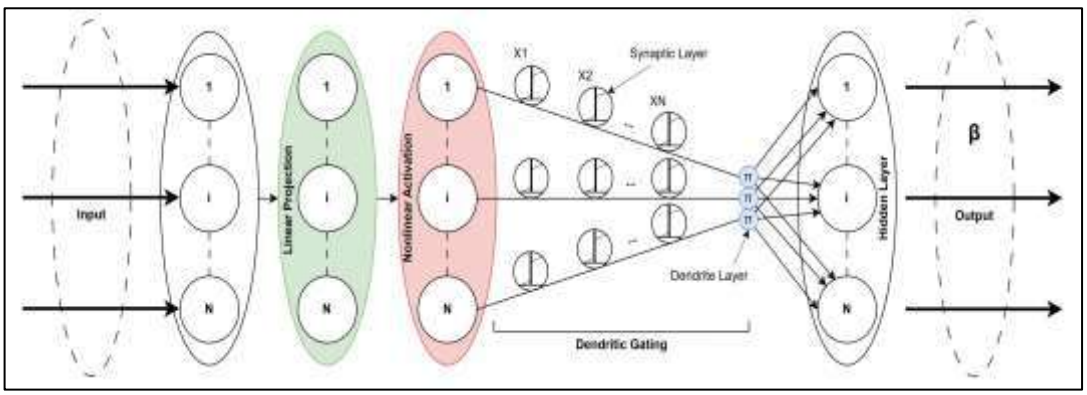


Figure 4: Architecture of the Proposed Method: Randomized Dendritic Neuron Model (R-DNM).
Source: Authors, (2026).

The forward pass process in R-DNM is identical to that of DNM up to the formation of the final hidden-layer representation H. Consequently, the computational stages directly follow the previously defined formulations, where the synapse and linear branch layers employ Formula (1)–Formula (3), the non-linear activation layer uses Formula (4), and the dendritic gating layer applies Formula (5)–Formula (7), such that the hidden representation $H \in R^{N \times n_{hidden}}$, for N data samples, is obtained through the same operations. The mapping to the output space also remains unchanged, namely the computation of the logits using Formula (8), followed by the calculation of the predictive probabilities using Formula (9), with numerical clamping applied. The fundamental difference lies in the optimization strategy. In R-DNM, since the hidden parameters are fixed, the hidden matrix H_{train} can be computed once through a forward pass without tracking gradients. The optimization process is therefore focused solely on β, by minimizing a regularized least-squares objective function (ridge regression), as defined in Formula (10):

$$\mathcal{J}(\beta) = \left|\left|H_{train}\beta - Y\right|\right|_F^2 + \lambda\left|\left|\beta\right|\right|_F^2 \tag{10}$$

Where $H_{train} \in R^{N_{train} \times n_{hidden}}$ denotes the hidden matrix obtained from the training data, $y_{train} \in \{1, \ldots, n_{classes}\}^{N_{train}}$ represents the target class labels for the training samples, $\lambda \geq 0$ is the ridge regularization hyperparameter, and $||\cdot||_F$ denotes the Frobenius norm, which is equivalent to $\sum_{i=1}^{N_{train}} \left|\left|(H_{train}\beta)_{i,:} - e_{y_i}\right|\right|_2^2 + \lambda \sum_{j,k} \beta_{j,k}^2$, where e_{y_i} denotes the canonical basis vector corresponding to the class label y_i. To obtain the optimal solution, the partial derivative of the objective function with respect to β is set to zero:

$$\frac{\partial \mathcal{J}}{\partial \beta} = 2H_{train}^{\top}(H_{train}\beta - Y) + 2\lambda\beta = 0 \tag{11}$$

Which yields

$$H_{train}^{\top}H_{train}\beta + \lambda\beta = H_{train}^{\top}Y \tag{12}$$

And can be rewritten as

$$(H_{train}^{\top}H_{train} + \lambda I)\beta = H_{train}^{\top}Y \tag{13}$$

Where $I \in R^{n_{hidden} \times n_{hidden}}$ is the identity matrix. The final closed-form solution is given in Formula (14):

$$\beta^* = (H_{train}^{\top}H_{train} + \lambda I)^{-1}H_{train}^{\top}Y \tag{14}$$

This approach allows the training of R-DNM to be completed in a single matrix computation step, with a computational complexity of $O(Nn_{hidden}^2 + n_{hidden}^3)$ due to the matrix inversion, which is substantially faster and more memory efficient than the iterative gradient-based optimization used in DNM. In addition, the randomization of hidden parameters provides an implicit ensemble effect that enhances robustness to noise and reduces overfitting. The complete training and inference procedure of the Randomized Dendritic Neuron Model (R-DNM) is summarized in Figure 5.

***Input**: Training data, validation data, n_hidden, λ (ridge), activation functions, random seed*
***Output**: Optimal output weights β* and validation metrics*
Start *1. Set random seed for reproducibility.* *2. Randomly initialize and fix all hidden-layer parameters (synaptic, linear branch, dendritic gating).* *3. Compute training hidden representation:* *- Perform full forward pass through synapse → linear branch → activation → dendritic gating.* *- Obtain hidden matrix H_train.* *4. Analytically solve ridge regression to compute optimal output weights β*.* *5. Inference on validation set:* *- Perform forward pass to obtain probabilities P_val.* *- Evaluate performance metrics against y_val.* ***End***

Figure 5: Training and Inference of the Randomized Dendritic Neuron Model (R-DNM).
Source: Authors, (2026).

II.4 EXPERIMENTAL SETUP AND EVALUATION

This study evaluates three neuromorphic models—the Dendritic Neuron Model (DNM), its multiclass extension Multi-Output Dendritic Neuron Model (MODN), and the proposed Randomized Dendritic Neuron Model (R-DNM)—alongside established baseline classifiers: Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). All experiments were conducted under identical conditions across four medical datasets to ensure fair and reproducible comparisons.

MODN extends DNM directly to multiclass classification by expanding the output weight matrix dimensionality while preserving the internal neuromorphic processing pipeline. In contrast, R-DNM retains the same dendritic-based hidden architecture up to the hidden representation but introduces a key innovation: hidden-layer parameters are randomly initialized and fixed, with only the output weights learned analytically via closed-form ridge regression. This design substantially reduces computational complexity and memory usage compared to the iterative backpropagation required by DNM and MODN, enabling higher efficiency without compromising biologically inspired dendritic non-linearities [53], [54].

Model performance was assessed using accuracy, precision, recall, F1-score, and log-loss to evaluate predictive capability, with particular emphasis on precision–recall balance for imbalanced datasets and probabilistic calibration. Computational efficiency was measured through training and inference time as well as memory consumption. Features were normalized using StandardScaler fitted exclusively on training data in each fold to prevent data leakage. All experiments were executed on a Windows 11 Home 64-bit system equipped with an AMD Ryzen 5 6600H processor (~3.3 GHz, 12 cores), 16 GB RAM, and 8,290 MB total graphics memory (483 MB dedicated, 7,806 MB shared). The complete experimental workflow is detailed in Figure 6.

***Input**: Dataset X, labels y, k = 10 folds*
***Output**: Average classification and computational metrics*
Start
1. Load dataset and labels.
2. Partition data into 10 folds.
3. For each fold:
- Split into training and test sets.
- Normalize features using StandardScaler (fit on training only).
- Train model (DNM, MODN, R-DNM, RF, SVM, or MLP).
- Perform inference on test set.
- Compute accuracy, precision, recall, F1-score, log-loss, computation time, and memory usage.
4. Average all metrics across the 10 folds.
5. Report final results.
End

Figure 6: Experimental and Evaluation Procedure.
Source: Authors, (2026).

III. RESULTS AND DISCUSSIONS

The performance evaluation of the Dendritic Neuron Model (DNM), Multi-Output Dendritic Neuron Model (MODN), and the proposed Randomized Dendritic Neuron Model (R-DNM) on the Pima Indians Diabetes dataset demonstrates the clear superiority of R-DNM in terms of computational efficiency and overall effectiveness. As summarized in Table 2, R-DNM consistently requires only 0.0020–0.0024 seconds for training and inference with memory usage of 0.0009 MB across all activation function combinations, in stark contrast to DNM and MODN, which require 0.4301–0.6597 seconds and 0.0063–0.0080 MB, respectively. This substantial reduction in computational cost is primarily due to the fixed random initialization of hidden-layer parameters—including synaptic, linear branch, and dendritic gating layers—drawn from a standard Gaussian distribution, with only the output weights optimized analytically using regularized least-squares, thereby reducing training complexity to a single dominant matrix operation.

In terms of classification performance, R-DNM outperforms DNM and MODN in nearly all activation configurations, except for the tanh–tanh setting, which yields the poorest results for all models due to severe saturation and vanishing gradients during iterative backpropagation. The optimal configuration for R-DNM is relu–tanh, achieving the highest average accuracy of 0.7552 ± 0.0362, precision of 0.7546 ± 0.0359, recall of 0.7552 ± 0.0362, F1-score of 0.7529 ± 0.0343, and log-loss of 0.5668 ± 0.0218. Relative to DNM and MODN, R-DNM exhibits consistent accuracy gains across several configurations, including +0.0169 for sigmoid–sigmoid, +0.0247 for sigmoid–tanh, and +0.0338 for relu–tanh, accompanied by reduced standard deviations that indicate improved cross-fold stability. These improvements are attributed to the implicit ensemble effect of large-scale random hidden representations and the preserved effectiveness of dendritic multiplicative gating for selective nonlinear modulation, particularly when ReLU is employed at the input layer, resulting in lower numerical instability and substantially reduced log-loss compared to DNM and MODN.

Table 2: Performance Comparison of Dendritic Neuron Models with Different Activation Function Combinations on Pima Indians Diabetes Dataset.

Activation Function (Input - Dendrite)	Method	Accuracy	Precision	Recall	F1-score	Log-loss	Computation time (s)	Total memory (MB)
sigmoid - sigmoid	DNM	0.7123 ± 0.0490	0.7109 ± 0.0514	0.7123 ± 0.0490	0.7093 ± 0.0486	1.3550 ± 0.3666	0.6382 ± 0.0664	0.0077 ± 0.0011
	MODN	0.7123 ± 0.0490	0.7109 ± 0.0514	0.7123 ± 0.0490	0.7093 ± 0.0486	1.3550 ± 0.3666	0.4548 ± 0.0403	0.0080 ± 0.0031
	Proposed Method: R-DNM	0.7292 ± 0.0264	0.7300 ± 0.0266	0.7292 ± 0.0264	0.7268 ± 0.0246	0.5621 ± 0.0243	0.0022 ± 0.0011	0.0009 ± 0.0000
sigmoid - tanh	DNM	0.7110 ± 0.0471	0.7099 ± 0.0468	0.7110 ± 0.0471	0.7087 ± 0.0468	1.7905 ± 0.3539	0.6232 ± 0.0813	0.0073 ± 0.0006
	MODN	0.7110 ± 0.0471	0.7099 ± 0.0468	0.7110 ± 0.0471	0.7087 ± 0.0468	1.7905 ± 0.3539	0.4871 ± 0.0622	0.0074 ± 0.0010

	Proposed Method: R-DNM	0.7357 ± 0.0488	0.7389 ± 0.0434	0.7357 ± 0.0488	0.7348 ± 0.0461	0.5619 ± 0.0204	0.0021 ± 0.0009	0.0009 ± 0.0002
sigmoid - relu	DNM	0.7251 ± 0.0437	0.7233 ± 0.0409	0.7251 ± 0.0437	0.7191 ± 0.0426	3.2714 ± 0.5157	0.6316 ± 0.0735	0.0072 ± 0.0004
	MODN	0.7251 ± 0.0437	0.7233 ± 0.0409	0.7251 ± 0.0437	0.7191 ± 0.0426	3.2714 ± 0.5157	0.4459 ± 0.0268	0.0066 ± 0.0003
	Proposed Method: R-DNM	0.7331 ± 0.0411	0.7377 ± 0.0399	0.7331 ± 0.0411	0.7325 ± 0.0385	0.5724 ± 0.0239	0.0021 ± 0.0004	0.0009 ± 0.0000
tanh - sigmoid	DNM	0.7161 ± 0.0353	0.7169 ± 0.0333	0.7161 ± 0.0353	0.7147 ± 0.0343	1.7314 ± 0.3129	0.6283 ± 0.0776	0.0072 ± 0.0005
	MODN	0.7161 ± 0.0353	0.7169 ± 0.0333	0.7161 ± 0.0353	0.7147 ± 0.0343	1.7314 ± 0.3129	0.4528 ± 0.0208	0.0067 ± 0.0005
	Proposed Method: R-DNM	0.7356 ± 0.0453	0.7431 ± 0.0444	0.7356 ± 0.0453	0.7362 ± 0.0436	0.5731 ± 0.0276	0.0024 ± 0.0011	0.0009 ± 0.0002
tanh - tanh	DNM	0.6408 ± 0.0503	0.6314 ± 0.0503	0.6408 ± 0.0503	0.6332 ± 0.0484	2.6561 ± 0.3843	0.6597 ± 0.0869	0.0072 ± 0.0004
	MODN	0.6408 ± 0.0503	0.6314 ± 0.0503	0.6408 ± 0.0503	0.6332 ± 0.0484	2.6561 ± 0.3843	0.4623 ± 0.0512	0.0071 ± 0.0003
	Proposed Method: R-DNM	0.6432 ± 0.0458	0.6228 ± 0.0552	0.6432 ± 0.0458	0.6243 ± 0.0479	0.6399 ± 0.0153	0.0020 ± 0.0009	0.0009 ± 0.0000
tanh - relu	DNM	0.7056 ± 0.0400	0.6992 ± 0.0392	0.7056 ± 0.0400	0.6979 ± 0.0391	4.1604 ± 0.5175	0.6390 ± 0.0618	0.0072 ± 0.0007
	MODN	0.7056 ± 0.0400	0.6992 ± 0.0392	0.7056 ± 0.0400	0.6979 ± 0.0391	4.1604 ± 0.5175	0.4472 ± 0.0209	0.0071 ± 0.0005
	Proposed Method: R-DNM	0.7136 ± 0.0424	0.7227 ± 0.0488	0.7136 ± 0.0424	0.7143 ± 0.0433	0.5817 ± 0.0249	0.0020 ± 0.0009	0.0009 ± 0.0000
relu - sigmoid	DNM	0.7200 ± 0.0420	0.7160 ± 0.0494	0.7200 ± 0.0420	0.7126 ± 0.0425	4.1586 ± 0.5363	0.6541 ± 0.0806	0.0070 ± 0.0003
	MODN	0.7200 ± 0.0420	0.7160 ± 0.0494	0.7200 ± 0.0420	0.7126 ± 0.0425	4.1586 ± 0.5363	0.4607 ± 0.0450	0.0068 ± 0.0005
	Proposed Method: R-DNM	0.7435 ± 0.0525	0.7412 ± 0.0552	0.7435 ± 0.0525	0.7391 ± 0.0528	0.5679 ± 0.0238	0.0021 ± 0.0004	0.0009 ± 0.0001
relu - tanh	DNM	0.7214 ± 0.0484	0.7126 ± 0.0583	0.7214 ± 0.0484	0.7053 ± 0.0587	4.3003 ± 0.8294	0.6267 ± 0.0480	0.0072 ± 0.0003
	MODN	0.7214 ± 0.0484	0.7126 ± 0.0583	0.7214 ± 0.0484	0.7053 ± 0.0587	4.3003 ± 0.8294	0.4301 ± 0.0144	0.0070 ± 0.0005
	Proposed Method: R-DNM	0.7552 ± 0.0362	0.7546 ± 0.0359	0.7552 ± 0.0362	0.7529 ± 0.0343	0.5668 ± 0.0218	0.0022 ± 0.0006	0.0009 ± 0.0000
relu - relu	DNM	0.6901 ± 0.0285	0.6770 ± 0.0402	0.6901 ± 0.0285	0.6617 ± 0.0372	4.9407 ± 0.4547	0.6260 ± 0.0588	0.0072 ± 0.0004
	MODN	0.6901 ± 0.0285	0.6770 ± 0.0402	0.6901 ± 0.0285	0.6617 ± 0.0372	4.9407 ± 0.4547	0.4440 ± 0.0382	0.0069 ± 0.0004
	Proposed Method: R-DNM	0.7306 ± 0.0494	0.7250 ± 0.0530	0.7306 ± 0.0494	0.7240 ± 0.0512	0.5856 ± 0.0220	0.0023 ± 0.0012	0.0009 ± 0.0000

Source: Authors, (2026).

The evaluation of the Dendritic Neuron Model (DNM), Multi-Output Dendritic Neuron Model (MODN), and the proposed Randomized Dendritic Neuron Model (R-DNM) on the Breast Cancer Wisconsin (Diagnostic) dataset consistently demonstrates the computational advantage of R-DNM across all activation function combinations. As reported in Table 3, R-DNM achieves training and inference times ranging from 0.0023 to 0.0041 seconds with a memory usage of 0.0009 MB, whereas DNM and MODN require substantially longer execution times (0.4444–0.8046 seconds) and larger memory footprints (0.0063–0.0077 MB). This efficiency gain is a direct consequence of fixing randomly initialized hidden-layer parameters and optimizing only the output weights analytically via ridge regression, thereby avoiding iterative backpropagation and gradient storage.

In contrast, classification performance on this dataset generally favors DNM and MODN due to the extremely high feature separability and low intrinsic complexity of the task. The best overall results are achieved under the sigmoid–sigmoid configuration, where DNM and MODN obtain an accuracy and F1-score of 0.9719 ± 0.0195 with a log-loss of 0.1806 ± 0.1761, while R-DNM reaches an accuracy of 0.9473 ± 0.0283 and a higher log-loss of 0.3881 ± 0.0255. Similar patterns are observed for sigmoid–tanh (0.9596 versus 0.9508) and sigmoid–relu (0.9684 versus 0.9456) configurations. Notably, R-DNM outperforms both DNM and MODN in activation settings susceptible to numerical instability, such as relu–relu and relu–tanh, where gradient-based models experience dead ReLU and exploding gradients, while R-DNM maintains stable accuracy above 0.9290 and consistently low log-loss due to its gradient-independent formulation.

Table 3: Performance Comparison of Dendritic Neuron Models with Different Activation Function Combinations on Breast Cancer Wisconsin (Diagnostic) Dataset.

Activation Function (Input - Dendrite)	Method	Accuracy	Precision	Recall	F1-score	Log-loss	Computation time (s)	Total memory (MB)
sigmoid - sigmoid	DNM	0.9719 ± 0.0195	0.9738 ± 0.0180	0.9719 ± 0.0195	0.9719 ± 0.0196	0.1806 ± 0.1761	0.8046 ± 0.0895	0.0077 ± 0.0035
	MODN	0.9719 ± 0.0195	0.9738 ± 0.0180	0.9719 ± 0.0195	0.9719 ± 0.0196	0.1806 ± 0.1761	0.4444 ± 0.0159	0.0075 ± 0.0028
	Proposed Method: R-DNM	0.9473 ± 0.0283	0.9506 ± 0.0262	0.9473 ± 0.0283	0.9467 ± 0.0287	0.3881 ± 0.0255	0.0028 ± 0.0013	0.0009 ± 0.0000
sigmoid - tanh	DNM	0.9596 ± 0.0261	0.9629 ± 0.0225	0.9596 ± 0.0261	0.9595 ± 0.0261	0.3295 ± 0.2548	0.7351 ± 0.0624	0.0065 ± 0.0003
	MODN	0.9596 ± 0.0261	0.9629 ± 0.0225	0.9596 ± 0.0261	0.9595 ± 0.0261	0.3295 ± 0.2548	0.4418 ± 0.0096	0.0065 ± 0.0005
	Proposed Method: R-DNM	0.9508 ± 0.0291	0.9524 ± 0.0282	0.9508 ± 0.0291	0.9503 ± 0.0296	0.3815 ± 0.0224	0.0041 ± 0.0028	0.0009 ± 0.0000
sigmoid - relu	DNM	0.9684 ± 0.0270	0.9717 ± 0.0225	0.9684 ± 0.0270	0.9683 ± 0.0269	0.3294 ± 0.2895	0.7323 ± 0.0898	0.0067 ± 0.0003
	MODN	0.9684 ± 0.0270	0.9717 ± 0.0225	0.9684 ± 0.0270	0.9683 ± 0.0269	0.3294 ± 0.2895	0.4593 ± 0.0332	0.0066 ± 0.0004
	Proposed Method: R-DNM	0.9456 ± 0.0318	0.9477 ± 0.0306	0.9456 ± 0.0318	0.9451 ± 0.0324	0.4021 ± 0.0269	0.0033 ± 0.0022	0.0009 ± 0.0002
tanh - sigmoid	DNM	0.9438 ± 0.0463	0.9479 ± 0.0422	0.9438 ± 0.0463	0.9439 ± 0.0458	0.3774 ± 0.3364	0.7307 ± 0.0825	0.0067 ± 0.0004
	MODN	0.9438 ± 0.0463	0.9479 ± 0.0422	0.9438 ± 0.0463	0.9439 ± 0.0458	0.3774 ± 0.3364	0.4509 ± 0.0399	0.0065 ± 0.0004
	Proposed Method: R-DNM	0.9456 ± 0.0253	0.9486 ± 0.0238	0.9456 ± 0.0253	0.9451 ± 0.0258	0.3926 ± 0.0195	0.0032 ± 0.0013	0.0009 ± 0.0000
tanh - tanh	DNM	0.7647 ± 0.0511	0.7673 ± 0.0496	0.7647 ± 0.0511	0.7618 ± 0.0505	1.4982 ± 0.5172	0.7462 ± 0.1371	0.0069 ± 0.0007
	MODN	0.7647 ± 0.0511	0.7673 ± 0.0496	0.7647 ± 0.0511	0.7618 ± 0.0505	1.4982 ± 0.5172	0.4587 ± 0.0402	0.0066 ± 0.0003
	Proposed Method: R-DNM	0.6750 ± 0.0633	0.6745 ± 0.0630	0.6750 ± 0.0633	0.6740 ± 0.0630	0.6279 ± 0.0314	0.0112 ± 0.0247	0.0009 ± 0.0000
tanh - relu	DNM	0.9298 ± 0.0323	0.9326 ± 0.0320	0.9298 ± 0.0323	0.9292 ± 0.0323	0.9764 ± 0.5727	0.7097 ± 0.0670	0.0066 ± 0.0003
	MODN	0.9298 ± 0.0323	0.9326 ± 0.0320	0.9298 ± 0.0323	0.9292 ± 0.0323	0.9764 ± 0.5727	0.4582 ± 0.0605	0.0065 ± 0.0003
	Proposed Method: R-DNM	0.9280 ± 0.0298	0.9324 ± 0.0253	0.9280 ± 0.0298	0.9278 ± 0.0299	0.4216 ± 0.0241	0.0032 ± 0.0015	0.0009 ± 0.0000
relu - sigmoid	DNM	0.9508 ± 0.0280	0.9533 ± 0.0262	0.9508 ± 0.0280	0.9503 ± 0.0286	0.7846 ± 0.4474	0.7179 ± 0.0733	0.0067 ± 0.0004
	MODN	0.9508 ± 0.0280	0.9533 ± 0.0262	0.9508 ± 0.0280	0.9503 ± 0.0286	0.7846 ± 0.4474	0.4660 ± 0.0120	0.0065 ± 0.0003
	Proposed Method: R-DNM	0.9509 ± 0.0383	0.9512 ± 0.0385	0.9509 ± 0.0383	0.9506 ± 0.0387	0.4086 ± 0.0182	0.0030 ± 0.0011	0.0009 ± 0.0000
relu - tanh	DNM	0.9139 ± 0.0336	0.9180 ± 0.0311	0.9139 ± 0.0336	0.9125 ± 0.0358	1.3294 ± 0.4790	0.7290 ± 0.0813	0.0067 ± 0.0003
	MODN	0.9139 ± 0.0336	0.9180 ± 0.0311	0.9139 ± 0.0336	0.9125 ± 0.0358	1.3294 ± 0.4790	0.4480 ± 0.0177	0.0067 ± 0.0006
	Proposed Method: R-DNM	0.9298 ± 0.0368	0.9321 ± 0.0352	0.9298 ± 0.0368	0.9299 ± 0.0364	0.4259 ± 0.0234	0.0027 ± 0.0007	0.0009 ± 0.0000
relu - relu	DNM	0.5553 ± 0.0699	0.5577 ± 0.0663	0.5553 ± 0.0699	0.5535 ± 0.0665	7.0892 ± 1.1151	0.7069 ± 0.0455	0.0065 ± 0.0004
	MODN	0.5553 ± 0.0699	0.5577 ± 0.0663	0.5553 ± 0.0699	0.5535 ± 0.0665	7.0892 ± 1.1151	0.4483 ± 0.0522	0.0063 ± 0.0005
	Proposed Method: R-DNM	0.9333 ± 0.0406	0.9363 ± 0.0379	0.9333 ± 0.0406	0.9325 ± 0.0415	0.4638 ± 0.0195	0.0023 ± 0.0010	0.0009 ± 0.0000

Source: Authors, (2026).

The performance assessment of the Dendritic Neuron Model (DNM), Multi-Output Dendritic Neuron Model (MODN), and the proposed Randomized Dendritic Neuron Model (R-DNM) on the large-scale and balanced Cardiovascular Disease dataset (approximately 70,000 samples with a near 50:50 class ratio) demonstrates the sustained computational advantage of R-DNM under high data volume.

As summarized in Table 4, R-DNM requires 1.8955–2.0362 seconds for training and inference with memory usage in the range of 0.0015–0.0016 MB, whereas DNM and MODN demand substantially longer execution times (7.6023–9.0608 seconds) and higher memory consumption, reaching up to 2.2401 ± 6.6981 MB for DNM and 0.0068–0.0080 MB for MODN. This efficiency is achieved through fixed random initialization of hidden-layer parameters and single-step analytical optimization of output weights via ridge regression, thereby avoiding the linear growth of computational cost associated with iterative backpropagation as dataset size increases. In terms of classification performance, results differ from those observed on smaller datasets, with DNM and MODN outperforming R-DNM across most activation configurations. The best overall performance is obtained under the sigmoid–sigmoid setting, where DNM and MODN achieve an accuracy of 0.7288 ± 0.0046, an F1-score of 0.7285 ± 0.0046, and a log-loss of 1.0231 ± 0.0168, compared to R-DNM's accuracy of 0.6919 ± 0.0071.

Similar trends are observed for sigmoid–tanh and tanh–sigmoid combinations, reflecting the superior capacity of end-to-end gradient-based optimization to exploit the dataset's balanced distribution and moderate-to-high complexity. Nevertheless, R-DNM demonstrates clear advantages in activation settings prone to gradient instability, such as relu-based configurations, where DNM and MODN experience severe performance degradation (e.g., accuracy of 0.6333 and log-loss of 5.8437 for relu–relu), while R-DNM maintains stable accuracy of 0.7081 and a substantially lower log-loss of 0.6159. Overall, these findings emphasize the trade-off between predictive accuracy and computational efficiency, positioning R-DNM as a scalable and robust alternative for large-scale medical applications with constrained computational resources.

Table 4: Performance Comparison of Dendritic Neuron Models with Different Activation Function Combinations on Cardiovascular Disease Dataset.

Activation Function (Input - Dendrite)	Method	Accuracy	Precision	Recall	F1-score	Log-loss	Computation time (s)	Total memory (MB)
sigmoid - sigmoid	DNM	0.7288 ± 0.0046	0.7299 ± 0.0047	0.7288 ± 0.0046	0.7285 ± 0.0046	1.0231 ± 0.0168	8.3109 ± 0.8423	2.2401 ± 6.6981
	MODN	0.7288 ± 0.0046	0.7299 ± 0.0047	0.7288 ± 0.0046	0.7285 ± 0.0046	1.0231 ± 0.0168	7.6841 ± 0.1648	0.0080 ± 0.0030
	Proposed Method: R-DNM	0.6919 ± 0.0071	0.6920 ± 0.0070	0.6919 ± 0.0071	0.6918 ± 0.0071	0.6216 ± 0.0022	1.9929 ± 0.1804	1.7116 ± 5.1303
sigmoid - tanh	DNM	0.7253 ± 0.0048	0.7259 ± 0.0049	0.7253 ± 0.0048	0.7251 ± 0.0048	1.4522 ± 0.0583	8.8255 ± 0.8411	0.0073 ± 0.0004
	MODN	0.7253 ± 0.0048	0.7259 ± 0.0049	0.7253 ± 0.0048	0.7251 ± 0.0048	1.4522 ± 0.0583	7.8948 ± 0.1335	0.0069 ± 0.0005
	Proposed Method: R-DNM	0.6856 ± 0.0063	0.6857 ± 0.0063	0.6856 ± 0.0063	0.6855 ± 0.0063	0.6242 ± 0.0021	2.0078 ± 0.2596	0.0016 ± 0.0002
sigmoid - relu	DNM	0.7252 ± 0.0061	0.7257 ± 0.0061	0.7252 ± 0.0061	0.7251 ± 0.0061	2.8883 ± 0.1165	8.4216 ± 0.8724	0.0068 ± 0.0005
	MODN	0.7252 ± 0.0061	0.7257 ± 0.0061	0.7252 ± 0.0061	0.7251 ± 0.0061	2.8883 ± 0.1165	7.6023 ± 0.2015	0.0072 ± 0.0002
	Proposed Method: R-DNM	0.6926 ± 0.0052	0.6926 ± 0.0051	0.6926 ± 0.0052	0.6925 ± 0.0052	0.6214 ± 0.0021	2.0014 ± 0.2312	0.0016 ± 0.0003
tanh - sigmoid	DNM	0.7223 ± 0.0047	0.7226 ± 0.0047	0.7223 ± 0.0047	0.7222 ± 0.0047	1.6475 ± 0.0496	8.7765 ± 0.8171	0.0070 ± 0.0005
	MODN	0.7223 ± 0.0047	0.7226 ± 0.0047	0.7223 ± 0.0047	0.7222 ± 0.0047	1.6475 ± 0.0496	8.1751 ± 0.8727	0.0072 ± 0.0003
	Proposed Method: R-DNM	0.6888 ± 0.0047	0.6888 ± 0.0047	0.6888 ± 0.0047	0.6887 ± 0.0047	0.6223 ± 0.0021	2.0293 ± 0.2706	0.0016 ± 0.0002
tanh - tanh	DNM	0.6815 ± 0.0062	0.6816 ± 0.0062	0.6815 ± 0.0062	0.6815 ± 0.0062	2.3898 ± 0.0770	9.0608 ± 0.7901	0.0071 ± 0.0005
	MODN	0.6815 ± 0.0062	0.6816 ± 0.0062	0.6815 ± 0.0062	0.6815 ± 0.0062	2.3898 ± 0.0770	8.2727 ± 0.4531	0.0070 ± 0.0005
	Proposed Method: R-DNM	0.6396 ± 0.0065	0.6396 ± 0.0065	0.6396 ± 0.0065	0.6396 ± 0.0065	0.6477 ± 0.0024	2.0362 ± 0.2474	0.0015 ± 0.0002
tanh - relu	DNM	0.6598 ± 0.0062	0.6599 ± 0.0062	0.6598 ± 0.0062	0.6598 ± 0.0062	4.8822 ± 0.1136	8.6689 ± 0.8216	0.0070 ± 0.0004
	MODN	0.6598 ± 0.0062	0.6599 ± 0.0062	0.6598 ± 0.0062	0.6598 ± 0.0062	4.8822 ± 0.1136	8.0335 ± 0.4503	0.0068 ± 0.0006
	Proposed Method: R-DNM	0.6868 ± 0.0057	0.6868 ± 0.0057	0.6868 ± 0.0057	0.6867 ± 0.0057	0.6239 ± 0.0018	1.9384 ± 0.0393	0.0016 ± 0.0002
relu - sigmoid	DNM	0.6714 ± 0.0082	0.6716 ± 0.0082	0.6714 ± 0.0082	0.6713 ± 0.0082	4.8688 ± 0.1219	8.5826 ± 0.9172	0.0072 ± 0.0004
	MODN	0.6714 ± 0.0082	0.6716 ± 0.0082	0.6714 ± 0.0082	0.6713 ± 0.0082	4.8688 ± 0.1219	7.6836 ± 0.2463	0.0072 ± 0.0005
	Proposed Method: R-DNM	0.7045 ± 0.0060	0.7047 ± 0.0060	0.7045 ± 0.0060	0.7045 ± 0.0060	0.6161 ± 0.0023	1.8955 ± 0.0417	0.0016 ± 0.0003
relu - tanh	DNM	0.6528 ± 0.0084	0.6534 ± 0.0083	0.6528 ± 0.0084	0.6524 ± 0.0085	5.2602 ± 0.1470	8.4964 ± 0.8339	0.0073 ± 0.0004
	MODN	0.6528 ± 0.0084	0.6534 ± 0.0083	0.6528 ± 0.0084	0.6524 ± 0.0085	5.2602 ± 0.1470	7.8824 ± 0.1661	0.0071 ± 0.0002
	Proposed Method: R-DNM	0.7012 ± 0.0062	0.7014 ± 0.0062	0.7012 ± 0.0062	0.7011 ± 0.0062	0.6178 ± 0.0024	1.9599 ± 0.1057	0.0016 ± 0.0003
relu - relu	DNM	0.6333 ± 0.0061	0.6358 ± 0.0068	0.6333 ± 0.0061	0.6316 ± 0.0064	5.8437 ± 0.0952	8.2332 ± 0.7251	0.0073 ± 0.0003
	MODN	0.6333 ± 0.0061	0.6358 ± 0.0068	0.6333 ± 0.0061	0.6316 ± 0.0064	5.8437 ± 0.0952	7.5852 ± 0.1776	0.0070 ± 0.0004
	Proposed Method: R-DNM	0.7081 ± 0.0072	0.7083 ± 0.0072	0.7081 ± 0.0072	0.7080 ± 0.0072	0.6159 ± 0.0030	1.9344 ± 0.2140	0.0015 ± 0.0002

Source: Authors, (2026).

To provide a concise cross-dataset comparison, Figures 7 and 8 summarize the average F1-score and computation time, averaged across all activation combinations, for DNM, MODN, and R-DNM. As illustrated in Figure 7, R-DNM achieves the highest average F1-score on the Pima Indians Diabetes dataset, attaining 0.7205 ± 0.0144 compared to 0.6958 ± 0.0149 for both DNM and MODN, reflecting the robustness introduced by implicit ensemble effects arising from random hidden representations on small and moderately imbalanced data. In contrast, DNM and MODN exhibit stronger average performance on the highly separable Breast Cancer Wisconsin dataset (0.9113 ± 0.0377 versus 0.8834 ± 0.0394 for R-DNM) and on the balanced Cardiovascular Disease dataset, where average F1-scores converge around 0.6886–0.6887, highlighting clear dataset-dependent performance trade-offs among the models.

Figure 8 further emphasizes the computational efficiency of R-DNM across all datasets. R-DNM consistently exhibits substantially lower average runtimes, ranging from 0.0020 to 0.0040 seconds on the smaller datasets and approximately 1.9800 seconds on the Cardiovascular Disease dataset. These observations are supported by paired t-tests: on the Pima Indians Diabetes dataset, R-DNM

demonstrates statistically significant improvements over DNM and MODN in F1-score ($t = -3.7695$, $p = 0.0014$, Cohen's $d = 1.6860$) and computation time ($t = 84.2454–110.6988$, $p < 0.0001$, $d = 37.6760–49.5060$).

On the Breast Cancer Wisconsin dataset, differences in F1-score are not statistically significant ($p = 0.1231$, $d = 0.7240$), whereas computation time differences remain highly significant ($p < 0.0001$, $d = 12.1830–17.3060$). Overall, these results confirm that R-DNM offers a favorable balance between predictive performance and computational efficiency, particularly for small-to-medium datasets and resource-constrained medical applications.

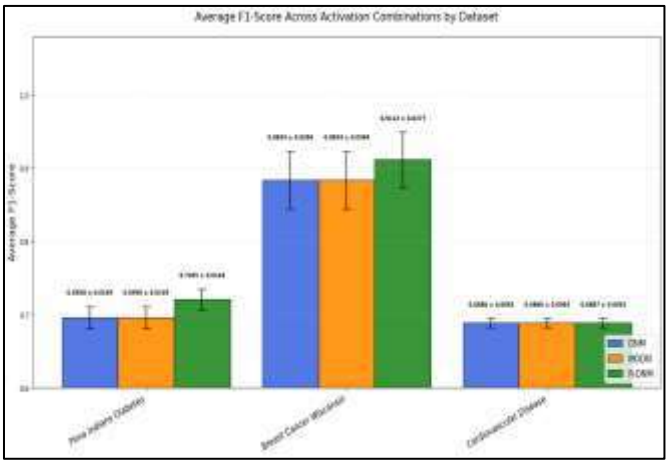


Figure 7: Average F1-Score Across Activation Combinations by Dataset.
Source: Authors, (2026).

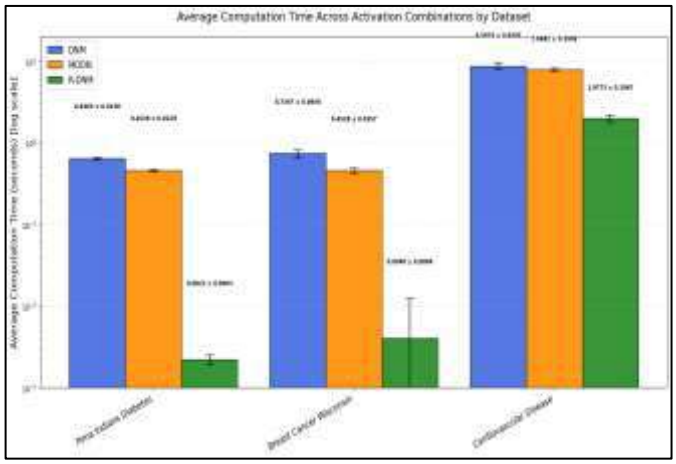


Figure 8: Average Computation Time Across Activation Combinations by Dataset.
Source: Authors, (2026).

The comparative evaluation of the best R-DNM configuration (ReLU–tanh activation) against established baseline classifiers—Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MLP)—on four medical datasets, as summarized in Table 5, indicates that R-DNM delivers competitive predictive performance while offering notable computational advantages. On the Pima Indians Diabetes dataset, R-DNM achieves an accuracy of 0.7552 ± 0.0362 and an F1-score of 0.7529 ± 0.0343, results that closely approach those of RF (0.7694 ± 0.0453) and SVM (0.7617 ± 0.0296) and are comparable to MLP (0.7525 ± 0.0435).

Similarly, on the Breast Cancer Wisconsin dataset, R-DNM attains an accuracy of 0.9333 ± 0.0406, slightly below the stronger baselines such as SVM and MLP (0.9754), while on the large-scale Cardiovascular Disease dataset it records an accuracy of 0.7081 ± 0.0072, marginally lower than RF (0.7251), SVM (0.7276), and MLP (0.7341). Despite these modest performance gaps, R-DNM consistently demonstrates favorable log-loss values in several cases and, more importantly, exhibits markedly superior efficiency, with training times ranging from 0.0022 to 1.9344 seconds and memory usage of approximately 0.0009–0.0015 MB, which is orders of magnitude lower than those required by RF (4.0143–39.6902 s), SVM (0.0548–1512.9478 s), and MLP (1.4114–43.9314 s).

Table 5: Comparison of Best R-DNM Configuration with Baseline Classifiers on Medical Datasets.

Dataset	Method	Accuracy	Precision	Recall	F1-score	Log-loss	Computation time (s)	Total memory (MB)
Pima Indians Diabetes	R-DNM (relu - tanh)	0.7552 ± 0.0362	0.7546 ± 0.0359	0.7552 ± 0.0362	0.7529 ± 0.0343	0.5668 ± 0.0218	0.0022 ± 0.0006	0.0009 ± 0.0000
	Random Forest	0.7694 ± 0.0453	0.7664 ± 0.0483	0.7694 ± 0.0453	0.7649 ± 0.0470	0.4857 ± 0.0573	4.0143 ± 0.4170	0.2591 ± 0.0046
	SVM	0.7617 ± 0.0296	0.7597 ± 0.0340	0.7617 ± 0.0296	0.7542 ± 0.0282	0.4936 ± 0.0455	0.1214 ± 0.0111	0.0368 ± 0.0008
	MLP	0.7525 ± 0.0435	0.7509 ± 0.0448	0.7525 ± 0.0435	0.7495 ± 0.0444	0.5760 ± 0.1116	5.0091 ± 0.3672	0.0593 ± 0.0013
Breast Cancer Wisconsin (Diagnostic)	R-DNM (relu - tanh)	0.9333 ± 0.0406	0.9363 ± 0.0379	0.9333 ± 0.0406	0.9325 ± 0.0415	0.4638 ± 0.0195	0.0023 ± 0.0010	0.0009 ± 0.0000
	Random Forest	0.9544 ± 0.0296	0.9565 ± 0.0279	0.9544 ± 0.0296	0.9541 ± 0.0301	0.1207 ± 0.0511	4.5808 ± 0.4896	0.2629 ± 0.0044
	SVM	0.9754 ± 0.0195	0.9767 ± 0.0183	0.9754 ± 0.0195	0.9754 ± 0.0195	0.0798 ± 0.0403	0.0548 ± 0.0079	0.0330 ± 0.0010
	MLP	0.9754 ± 0.0195	0.9766 ± 0.0179	0.9754 ± 0.0195	0.9752 ± 0.0199	0.0868 ± 0.0712	1.4114 ± 0.1809	0.0691 ± 0.0010
Cardiovascular Disease	R-DNM (relu - tanh)	0.7081 ± 0.0072	0.7083 ± 0.0072	0.7081 ± 0.0072	0.7080 ± 0.0072	0.6159 ± 0.0030	1.9344 ± 0.2140	0.0015 ± 0.0002
	Random Forest	0.7251 ± 0.0044	0.7256 ± 0.0043	0.7251 ± 0.0044	0.7249 ± 0.0044	0.5598 ± 0.0043	39.6902 ± 1.4447	0.2681 ± 0.0092
	SVM	0.7276 ± 0.0052	0.7285 ± 0.0053	0.7276 ± 0.0052	0.7273 ± 0.0052	0.5656 ± 0.0041	1512.9478 ± 251.9854	4.8254 ± 0.0067

	MLP	0.7341 ± 0.0046	0.7349 ± 0.0044	0.7341 ± 0.0046	0.7339 ± 0.0047	0.5450 ± 0.0042	43.9314 ± 6.8151	0.0354 ± 0.0017

Source: Authors, (2026).

To further contextualize the performance of the proposed Randomized Dendritic Neuron Model (R-DNM), Table 6 compares it with recent state-of-the-art methods evaluated on the same benchmark datasets. On the Pima Indians Diabetes Database (PIDD), R-DNM with the optimal configuration (ReLU activation at the input layer and tanh at the dendritic gating layer) achieves an accuracy of 0.7552 and an F1-score of 0.7529, which are competitive with ensemble-based approaches reported by (accuracy 0.7710, F1-score 0.6900) [16] and (accuracy 0.7700, F1-score 0.7100) [55]. Importantly, R-DNM exhibits significantly higher computational efficiency, requiring only about 0.0022 seconds for training and inference, whereas ensemble methods incur substantially higher costs due to iterative optimizatio and meta-learner training. This efficiency advantage makes R-DNM particularly suitable for resource-constrained, real-time applications on small to moderately imbalanced datasets such as PIDD. On the Breast Cancer Wisconsin (Diagnostic) dataset, R-DNM attains an accuracy of 0.9333 and an F1-score of 0.9325, slightly below the Intelligent Dendritic Neural Model optimized with Biogeography-Based Optimization (BBO) by (accuracy 0.9594) [29], a difference attributable to the dataset's high feature separability where evolutionary optimization yields marginal gains.

Nevertheless, R-DNM completes training and inference in approximately 0.0023 seconds, far faster than optimization-based approaches. Similarly, on the large-scale and balanced Cardiovascular Disease dataset, R-DNM achieves an accuracy of 0.7081 and an F1-score of 0.7080, approaching the performance of the attention-based Feature Decomposition Deep Learning (FDDL) model by (accuracy 0.7552, F1-score 0.7522) [18], while requiring only about 1.9344 seconds for training and inference. Overall, Table 6 highlights that R-DNM offers a strong performance–efficiency trade-off, delivering competitive predictive results with orders-of-magnitude improvements in computational speed and memory efficiency, making it a promising solution for real-world medical applications in resource-limited clinical settings.

Table 6: Comparison of the Proposed R-DNM (Optimal Configuration: ReLU–tanh) with Recent State-of-the-Art Methods on Benchmark Medical Datasets.

Dataset	Author(s)	Method(s)	Accuracy	F1-score
Pima Indians Diabetes	[16]	Ensemble learning (Decision Tree, Random Forest, Support Vector Machine, Neural Network, Logistic Regression)	0.7710	0.6900
	[55]	Ensemble models (Random Forest, Gradient Boosting, Multilayer Perceptron, XGBoost)	0.7700	0.7100
	This Study	Randomized Dendritic Neuron Model (ReLU activation at input layer – Tanh activation at dendritic layer)	0.7552	0.7529
Breast Cancer Wisconsin (Diagnostic)	[29]	Optimized Dendritic Neuron Model using Biogeography-Based Optimization	0.9594	-
	This Study	Randomized Dendritic Neuron Model (ReLU activation at input layer – Tanh activation at dendritic layer)	0.9333	0.9325
Cardiovascular Disease	[18]	Feature Decomposition-Based Deep Learning (FDDL) with attention mechanism and SHAP	0.7552	0.7522
	This Study	Randomized Dendritic Neuron Model (ReLU activation at input layer – Tanh activation at dendritic layer)	0.7081	0.7080

Source: Authors, (2026).

These results reveal a clear performance-efficiency trade-off modulated by dataset characteristics. On smaller or moderately imbalanced datasets such as PIDD, R-DNM delivers predictive accuracy and F1-scores competitive with both conventional baselines (RF, SVM, MLP) and recent state-of-the-art ensemble methods, benefiting from the implicit ensemble effect of randomized hidden representations and the robustness of dendritic gating. On datasets with high feature separability (Breast Cancer Wisconsin) or large-scale balanced data (Cardiovascular Disease), iterative optimization in end-to-end models and advanced deep architectures tends to yield marginally superior discrimination; however, R-DNM consistently offers orders-of-magnitude improvements in computational speed and memory usage, while demonstrating greater resilience to gradient instabilities (e.g., in ReLU-based configurations).

This positions R-DNM as a highly practical alternative for real-world deployment in resource-constrained clinical settings. The proposed Randomized Dendritic Neuron Model (R-DNM) preserves the biologically inspired non-linear multiplicative gating of dendritic structures while achieving extreme computational efficiency through fixed random hidden parameters and closed-form output optimization. Compared to both conventional classifiers and state-of-the-art approaches reported in recent literature, R-DNM provides competitive predictive performance across diverse medical tasks, coupled with training and inference speeds that are substantially faster—often by two to three orders of magnitude—making it particularly suitable for resource-limited environments.

Despite these advances, limitations remain, including slightly reduced accuracy on datasets with exceptionally high separability or large-scale balanced distributions, as well as limited exploration of activation functions and ridge regularization hyperparameters. Future work should explore hybrid training strategies that combine randomization with selective fine-tuning of hidden parameters, integration with multi-modal clinical data, external validation on real-world cohorts, and potential hardware implementation on neuromorphic platforms to further improve scalability and energy efficiency in digital health applications.

IV. CONCLUSIONS

This study introduces the Randomized Dendritic Neuron Model (R-DNM), a novel variant of the Dendritic Neuron Model that preserves the biologically inspired neuromorphic architecture, including the multiplicative gating mechanism in the dendritic layer, while achieving substantial improvements in computational efficiency through randomization and fixation of hidden-layer parameters combined with closed-form analytical training of the output weights via ridge regression. The proposed model delivers competitive predictive performance across benchmark medical datasets—Pima Indians Diabetes Database, Breast Cancer Wisconsin (Diagnostic), and

Cardiovascular Disease—often approaching or matching recent state-of-the-art ensemble, optimized dendritic, and attention-enhanced deep learning methods, while providing training and inference speeds that are orders of magnitude faster and memory requirements that are significantly lower.

These characteristics position R-DNM as a practical and scalable solution for early detection of non-communicable diseases in resource-constrained clinical environments. Nonetheless, limitations include modestly reduced accuracy on datasets exhibiting exceptionally high feature separability or large-scale balanced distributions, as well as limited exploration of activation function combinations and regularization hyperparameters. Future research directions should include the development of hybrid training strategies that integrate randomization with selective fine-tuning, extension to multi-modal clinical data, rigorous external validation using real-world patient cohorts, and hardware implementation on neuromorphic platforms to further advance scalability and energy efficiency in digital health applications.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Michael Angello Qadosy Riyadi and Adinda Mariasti Dewi.
Methodology: Michael Angello Qadosy Riyadi.
Investigation: Michael Angello Qadosy Riyadi and Adinda Mariasti Dewi.
Discussion of results: Michael Angello Qadosy Riyadi and Adinda Mariasti Dewi.
Writing – Original Draft: Adinda Mariasti Dewi.
Writing – Review and Editing: Michael Angello Qadosy Riyadi and Adinda Mariasti Dewi.
Resources: Michael Angello Qadosy Riyadi and Adinda Mariasti Dewi.
Supervision: Michael Angello Qadosy Riyadi and Adinda Mariasti Dewi.
Approval of the final text: Michael Angello Qadosy Riyadi and Adinda Mariasti Dewi.

VI. REFERENCES

[1] R. Raj, R. Garg, and V. Kaur, "Data-Driven Insights into Cardiovascular Disease Mortality: A Global Analysis of Trends and Risks," in Proceedings of the Second International Conference on Cognitive Robotics and Intelligent Systems (ICC-ROBINS), Coimbatore, India, 2025, pp. 516–522. doi: 10.1109/ICC-ROBINS64345.2025.11086182.

[2] P. Saeedi et al., "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition," Diabetes Research and Clinical Practice, vol. 157, p. 107843, 2019, doi: 10.1016/j.diabres.2019.107843.

[3] G. A. Roth et al., "Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study," Journal of the American College of Cardiology, vol. 76, no. 25, pp. 2982–3021, 2020, doi: 10.1016/j.jacc.2020.11.010.

[4] A. N. Giaquinto et al., "Breast Cancer Statistics, 2022," CA: A Cancer Journal for Clinicians, vol. 72, no. 6, pp. 524–541, 2022, doi: 10.3322/caac.21754.

[5] Y. Zhang et al., "Global burden of female breast cancer: new estimates in 2022, temporal trend and future projections up to 2050 based on the latest release from GLOBOCAN," Journal of the National Cancer Center, vol. 5, no. 3, pp. 287–296, 2025, doi: 10.1016/j.jncc.2025.02.002.

[6] H. Sung et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[7] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature Medicine, vol. 25, no. 1, pp. 44–56, 2019, doi: 10.1038/s41591-018-0300-7.

[8] K. Khunti et al., "Diabetes and Multiple Long-term Conditions: A Review of Our Current Global Health Challenge," Diabetes Care, vol. 46, no. 12, pp. 2092–2101, 2023, doi: 10.2337/dci23-0035.

[9] K. Bhavsar, A. Abugabah, J. Singla, A. Ali, A. Bashir, and Nikita, "A Comprehensive Review on Medical Diagnosis Using Machine Learning," Computers, Materials & Continua, vol. 67, pp. 1997–2014, Dec. 2020, doi: 10.32604/cmc.2021.014943.

[10] A. Garg and V. Mago, "Role of machine learning in medical research: A survey," Computer Science Review, vol. 40, p. 100370, May 2021, doi: 10.1016/j.cosrev.2021.100370.

[11] S. Athisayamani, S. T, A. R. Singh, and others, "A novel double machine learning approach for detecting early breast cancer using advanced feature selection and dimensionality reduction techniques," Scientific Reports, vol. 15, p. 22971, 2025, doi: 10.1038/s41598-025-06426-7.

[12] M. Z. Awan, M. Zain, and K. Abodayeh, "Comparative analysis of machine learning models for breast cancer prediction and diagnosis: A dual-dataset approach," Indonesian Journal of Electrical Engineering and Computer Science, vol. Vol.34, pp. 2032–2044, Jun. 2024, doi: 10.11591/ijeecs.v34.i3.pp2032-2044.

[13] A. F. Ashour, M. M. Fouda, Z. M. Fadlullah, and M. I. Ibrahem, "Optimized Neural Networks for Diabetes Classification Using Pima Indians Diabetes Database," in 2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI), Mt Pleasant, MI, USA, 2024, pp. 1–7. doi: 10.1109/ICMI60790.2024.10585703.

[14] P. V. Sankar Ganesh and P. Sripriya, "A Comparative Review of Prediction Methods for Pima Indians Diabetes Dataset," in Computational Vision and Bio-Inspired Computing, S. Smys, J. M. R. S. Tavares, V. E. Balas, and A. M. Iliyasu, Eds., Cham: Springer International Publishing, 2020, pp. 735–750. doi: https://doi.org/10.1007/978-3-030-37218-7_83.

[15] T. J. H. Mim, Md. M. Hassan, B. K. Paul, and others, "Machine Learning Approaches for Cardiovascular Disease Prediction: A Comparative Study," Biomedical Materials & Devices, 2025, doi: 10.1007/s44174-025-00564-2.

[16] M. S. Reza, R. Amin, R. Yasmin, W. Kulsum, and S. Ruhi, "Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data," Heliyon, vol. 10, no. 2, p. e24536, 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e24536.

[17] K. A. Ahmed et al., "Advancing breast cancer prediction: Comparative analysis of ML models and deep learning-based multi-model ensembles on original and synthetic datasets," PLOS ONE, vol. 20, no. 6, p. e0326221, 2025, doi: 10.1371/journal.pone.0326221.

[18] L. Yu et al., "Interpretable machine learning model for cardiovascular disease risk prediction: a feature decomposition-based study," BMC Public Health, vol. 25, no. 1, p. 3639, 2025, doi: 10.1186/s12889-025-24921-4.

[19] I. Araf, A. Idri, and I. Chairi, "Cost-sensitive learning for imbalanced medical data: a review," Artificial Intelligence Review, vol. 57, p. 80, 2024, doi: 10.1007/s10462-023-10652-8.

[20] S. A. Alowais et al., "Revolutionizing healthcare: the role of artificial intelligence in clinical practice," BMC Medical Education, vol. 23, no. 1, p. 689, 2023, doi: 10.1186/s12909-023-04698-z.

[21] X. Chen, H. Fan, W. Chen, Y. Zhang, D. Zhu, and S. Song, "Explaining a Logic Dendritic Neuron Model by Using the Morphology of Decision Trees," Electronics, vol. 13, no. 19, 2024, doi: 10.3390/electronics13193911.

[22] S. Gao, M. Zhou, Y. Wang, J. Cheng, H. Yachi, and J. Wang, "Dendritic Neuron Model With Effective Learning Algorithms for Classification, Approximation, and Prediction," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 2, pp. 601–614, 2019, doi: 10.1109/TNNLS.2018.2846646.

[23] J. Ji, C. Tang, J. Zhao, T. Zheng, and Y. Todo, "A Survey on Dendritic Neuron Model: Mechanisms, Algorithms and Practical Applications," Neurocomputing, vol. 489, Mar. 2022, doi: 10.1016/j.neucom.2021.08.153.

[24] V. Takács et al., "Synaptic and dendritic architecture of different types of hippocampal somatostatin interneurons," PLOS Biology, vol. 22, no. 3, p. e3002539, 2024, doi: 10.1371/journal.pbio.3002539.

[25] Z. Wang, N. Yu, and F. Essaf, "A soma-synapses neuron model and FPGA implementation," Concurrency and Computation: Practice and Experience, vol. 35, no. 27, p. e7851, 2023, doi: 10.1002/cpe.7851.

[26] Y. Zhang, Y. Yang, X. Li, Z. Yuan, Y. Todo, and H. Yang, "A Dendritic Neuron Model Optimized by Meta-Heuristics with a Power-Law-Distributed Population Interaction Network for Financial Time-Series Forecasting," Mathematics, vol. 11, no. 5, p. 1251, 2023, doi: 10.3390/math11051251.

[27] Y. Yang, X. Li, H. Li, C. Zhang, Y. Todo, and H. Yang, "Yet Another Effective Dendritic Neuron Model Based on the Activity of Excitation and Inhibition," Mathematics, vol. 11, no. 7, p. 1701, 2023, doi: 10.3390/math11071701.

[28] C. Tang et al., "Visual Analytics of Learning Behavior Based on the Dendritic Neuron Model," in Knowledge Science, Engineering and Management, C. Cao, H. Chen, L. Zhao, J. Arshad, T. Asyhari, and Y. Wang, Eds., Singapore: Springer Nature Singapore, 2024, pp. 192–203. doi: https://doi.org/10.1007/978-981-97-5495-3_14.

[29] W. Xu, D. Jia, Z. Zhong, C. Li, and Z. Xu, "Intelligent Dendritic Neural Model for Classification Problems," Symmetry, vol. 14, no. 1, p. 11, 2022, doi: 10.3390/sym14010011.

[30] X. Luo and others, "Pruning method for dendritic neuron model based on dendrite layer significance constraints," CAAI Transactions on Intelligence Technology, vol. 8, no. 2, pp. 308–318, 2023, doi: 10.1049/cit2.12234.

[31] D. Jia, W. Xu, D. Liu, Z. Xu, Z. Zhong, and X. Ban, "Verification of Classification Model and Dendritic Neuron Model Based on Machine Learning," Discrete Dynamics in Nature and Society, vol. 2022, no. 1, p. 3259222, 2022, doi: https://doi.org/10.1155/2022/3259222.

[32] Y. M. Thant, S. Manzhos, M. Ihara, and M. Nukunudompanich, "On the Sufficiency of a Single Hidden Layer in Feed-Forward Neural Networks Used for Machine Learning of Materials Properties," Physchem, vol. 5, no. 1, p. 4, 2025, doi: 10.3390/physchem5010004.

[33] C. Peralez-González, J. Pérez-Rodríguez, and A. M. Durán-Rosal, "Boosting ridge for the extreme learning machine globally optimised for classification and regression problems," Scientific Reports, vol. 13, p. 11809, 2023, doi: 10.1038/s41598-023-38948-3.

[34] M. Toğaçar, B. Ergen, and Z. Cömert, "Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders," Medical Hypotheses, vol. 135, p. 109503, 2020, doi: 10.1016/j.mehy.2019.109503.

[35] J. Wang, F. Xie, F. Nie, and X. Li, "Generalized and Robust Least Squares Regression," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 5, pp. 7006–7020, May 2024, doi: 10.1109/TNNLS.2022.3213594.

[36] I. Perfilieva, N. Madrid, M. Ojeda-Aciego, P. Artiemjew, and A. Niemczynowicz, "A critical analysis of the theoretical framework of the Extreme Learning Machine," Neurocomputing, vol. 621, p. 129298, 2025, doi: https://doi.org/10.1016/j.neucom.2024.129298.

[37] D. Chellappan and H. Rajaguru, "Generalizability of machine learning models for diabetes detection a study with nordic islet transplant and PIMA datasets," Scientific Reports, vol. 15, p. 4479, 2025, doi: 10.1038/s41598-025-87471-0.

[38] A. I. Jony and A. K. B. Arnob, "Deep Learning Paradigms for Breast Cancer Diagnosis: A Comparative Study on Wisconsin Diagnostic Dataset," Malaysian Journal of Science and Advanced Technology, vol. 4, no. 2, pp. 109–117, Mar. 2024, doi: https://doi.org/10.56532/mjsat.v4i2.245.

[39] M. Dorraki et al., "Improving Cardiovascular Disease Prediction With Machine Learning Using Mental Health Data: A Prospective UK Biobank Study," JACC. Advances, vol. 3, no. 9, p. 101180, 2024, doi: 10.1016/j.jacadv.2024.101180.

[40] X. Pei et al., "Robustness of machine learning to color, size change, normalization, and image enhancement on micrograph datasets with large sample differences," Materials & Design, vol. 232, p. 112086, 2023, doi: https://doi.org/10.1016/j.matdes.2023.112086.

[41] V. Lumumba, D. Sang, G. Njoka, D. Musyimi, and Kavita, "Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models," American Journal of Theoretical and Applied Statistics, vol. 13, pp. 127–137, Oct. 2024, doi: 10.11648/j.ajtas.20241305.13.

[42] G. Naidu, T. Zuva, and E. M. Sibanda, "A Review of Evaluation Metrics in Machine Learning Algorithms," in Artificial Intelligence Application in Networks and Systems, R. Silhavy and P. Silhavy, Eds., Cham: Springer International Publishing, 2023, pp. 15–25. doi: https://doi.org/10.1007/978-3-031-35314-7_2.

[43] S. Paul, P. Ranjan, S. Kumar, and A. Kumar, "Disease Predictor Using Random Forest Classifier," in 2022 International Conference for Advancement in Technology (ICONAT), Goa, India, 2022, pp. 1–4. doi: 10.1109/ICONAT53423.2022.9726023.

[44] F. M. Megahed, Y. J. Chen, L. A. Jones-Farmer, and others, "Comparing classifier performance with baselines," Nature Methods, vol. 21, pp. 546–548, 2024, doi: 10.1038/s41592-024-02234-5.

[45] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," Information, vol. 15, no. 4, p. 235, 2024, doi: 10.3390/info15040235.

[46] T. Bikku, "Multi-layered deep learning perceptron approach for health risk prediction," Journal of Big Data, vol. 7, p. 50, 2020, doi: 10.1186/s40537-020-00316-7.

[47] W. Xu et al., "A Dendritic Neuron Model for Disease Prediction," in 2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), 2021, pp. 1118–1122. doi: 10.1109/AEECA52519.2021.9574192.

[48] C. Tang, J. Ji, Y. Todo, A. Shimada, W. Ding, and A. Hirata, "Dendritic Neural Network: A Novel Extension of Dendritic Neuron Model," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 8, no. 3, pp. 2228–2239, 2024, doi: 10.1109/TETCI.2024.3367819.

[49] R. Giryes, G. Sapiro, and A. M. Bronstein, "Corrections to 'Deep Neural Networks With Random Gaussian Weights: A Universal Classification Strategy?,'" IEEE Transactions on Signal Processing, vol. 68, pp. 529–531, 2020, doi: 10.1109/TSP.2019.2961228.

[50] T. Uelwer, T. Hoffmann, and S. Harmeling, "Non-iterative Phase Retrieval with Cascaded Neural Networks," in Artificial Neural Networks and Machine Learning – ICANN 2021, vol. 12892, I. Farkaš, P. Masulli, S. Otte, and S. Wermter, Eds., in Lecture Notes in Computer Science, vol. 12892. , Cham: Springer, 2021. doi: 10.1007/978-3-030-86340-1_24.

[51] M. Uzair and N. Jamil, "Effects of Hidden Layers on the Efficiency of Neural Networks," in 2020 IEEE 23rd International Multitopic Conference (INMIC), Bahawalpur, Pakistan, 2020, pp. 1–6. doi: 10.1109/INMIC50486.2020.9318195.

[52] D. Zhang, Z. Wang, and T. Masayoshi, "Neural-Network-Based Iterative Learning Control for Multiple Tasks," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 9, pp. 4178–4190, 2021, doi: 10.1109/TNNLS.2020.3017158.

[53] X. Wen, M. Zhou, A. Albeshri, L. Huang, X. Luo, and D. Ning, "Improving Classification Performance in Dendritic Neuron Models through Practical Initialization Strategies," Sensors, vol. 24, no. 6, p. 1729, 2024, doi: 10.3390/s24061729.

[54] Y. Ding, J. Yu, C. Gu, S. Gao, and C. Zhang, "A multi-in and multi-out dendritic neuron model and its optimization," Knowledge-Based Systems, vol. 286, p. 111442, 2024, doi: https://doi.org/10.1016/j.knosys.2024.111442.

[55] H. Zhou, S. Rahman, M. Angelova, C. R. Bruce, and C. Karmakar, "A robust and generalized framework in diabetes classification across heterogeneous environments," Computers in Biology and Medicine, vol. 186, p. 109720, 2025, doi: https://doi.org/10.1016/j.compbiomed.2025.109720.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1162-1169. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3299

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

HIGH-ACCURACY BINARY CLASSIFICATION OF DIABETIC RETINOPATHY FROM FUNDUS IMAGES USING CNN ARCHITECTURE

Forhad Javed[1], Ashraful Islam[2], Mohammed Hasibul Hasan Chowdhury[3], Sayma Sultana[4], Nazmul Islam Frahim[5], Atikur Rahman[6]

[1, 2, 3, 4, 5, 6]Department of Electrical and Electronic Engineering, Bangladesh Army International University of Science & Technology (BAIUST), Cumilla-3501, Bangladesh.

[1]https://orcid.org/0009-0003-4820-5726, [2]https://orcid.org/0000-0002-2406-1938, [3]https://orcid.org/0009-0006-8629-5787, [4]https://orcid.org/0009-0001-6001-1452, [5]https://orcid.org/0009-0006-5559-1292, [6]https://orcid.org/0009-0005-5799-8486

Email: jabedforhad6@gmail.com, ashrafulislam.eece@gmail.com, hasibul.jobs.007@gmail.com, supornasayma01@gmail.com, nazmulislam281st@gmail.com, atiksarkar113@gmail.com

ARTICLE INFO

Article History
Received: January 13, 2025
Revised: January 20, 2026
Accepted: January 30, 2026
Published: February 28, 2026

Keywords:
Diabetic retinopathy,
Diabetes mellitus,
Artificial intelligence,
Convolutional neural network.

ABSTRACT

Diabetic Retinopathy (DR), a retinal condition associated with diabetes mellitus, is a leading global cause of blindness. Early detection and treatment is crucial to prevent or mitigate vision loss. Researchers have developed various artificial intelligence-based methods to detect diabetic retinopathy from fundus images, aiming to enhance diagnostic accuracy and efficiency, thereby improving patient outcomes. In this paper, we have introduced a custom convolutional neural network (CNN) architecture for detecting diabetic retinopathy. We have used an open-source "Diabetic Retinopathy 224×224 Gaussian Filtered" dataset from Kaggle, comprising of 3,668 images. The images are then labelled into two classes where one class represents "No Diabetic Retinopathy (No_DR)" and the other class represents "Diabetic Retinopathy (DR)". Then we have applied min-max normalization to scale pixel values to a specified range [0, 1]. Afterwards, the dataset is divided into three parts to train, validate and test the proposed CNN model. Our method has achieved a testing accuracy of 95.27% and F1 score of 95.29%. Then we have performed 5-fold cross validation to observe the performance of the proposed automatic DR detection method. We have also compared our proposed method with some pre-trained models and some existing methods. In observation, it has been shown that the proposed technique outperformed many other existing techniques.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Diabetic eye diseases refer to a group of eye problems that can occur as a result of diabetes. Diabetic Retinopathy (DR) is among the most common diabetic eye diseases. It affects a significant portion of individuals living with diabetes, particularly those with poorly controlled blood sugar levels over an extended period of time. DR is caused as a result of Diabetes Mellitus that causes abnormal new blood vessels to grow on the surface of the retina [1]. As a result, the blood vessels of the retina may swell and/or leak fluid or blood. DR may not cause any harm in vision at first but over time it can get worse. It is very important to diagnose DR at the early stages.

But if DR is undiagnosed, it becomes more and more severe over the time through formation of Micro aneurysms, Exudates, Hemorrhages, Cotton Wool Spots, abnormal structure of the Optic Disc, abnormal Foveal Avascular Zone, Neovascularization's, Intra Retinal Microvascular Abnormalities and many more [1]. A person might see spots floating in their vision or may notice a general blurring of vision. Eventually that will result in severe vision loss and even blindness. According to the statistics of International Diabetes Federation [2], there are 537 million people suffering from diabetes globally in 2021 and it is expected to be 643 million and 783 million by 2030, and by 2045 respectively. It is also mentioned there that, by 2030, of all the diabetes patients, around 191 million people will suffer from DR and among them 56.3 million will be facing vision-threatening diabetic retinopathy [3].

According to a report [4] published in 2023, the number of people affected by diabetes in Bangladesh is more than 13 million. Out of them, around 1.85 million is suffering from diabetic retinopathy. This report also states that, a hospital based cross-sectional study was performed among 489 patients having type-2 diabetes, and the researchers have found that 18.8% of those patients were suffering from diabetic retinopathy. So, it can be said from the above discussions that the number of DR patients are increasing day by day. So, early diagnosis of diabetic retinopathy is extremely necessary for the diabetes patient. Artificial intelligence and machine learning techniques are used for the development of computer-aided diagnosis systems for the detection of diabetic retinopathy. These systems can automatically analyze fundus images and provide risk assessments or early detection of DR with high accuracy. Different pre-trained algorithms like VGG-16 [5], Xception [6], EfficientNet [7], ResNet [8] etc. are used for the automatic detection of DR.

The first triumph was taken in 1990s to classify retinal images in order to detect whether an eye has DR or not [9]. In this research, Gardner and his fellow researchers have proposed an artificial neural network (ANN) based tool that detects retinal images with DR. Their proposed tool provides 88.4% sensitivity and 83.5% specificity. In [10], the researchers have applied different algorithms i.e. random forest classifier, support vector machine, K-nearest neighbor, logistic regression, regression tree classifier, and the Naïve Bayes theorem to detect DR. They have found that, in terms of performance in evaluation and accuracy, the Naïve Bayes performs better among the six. According to [5] have used VGG-16 architecture for the automatic detection of diabetic retinopathy. They have used a dataset containing 3668 retinal images, captured through fundus photography. The model was then trained for 30 epochs. They have used ADAM optimizer to adjust the characteristics of the neural network. Categorical cross entropy loss function was used in their proposed method. The authors of [11] have proposed Inception-v3 based diabetic retinopathy detection process.

The datasets used in their work are from Kaggle, named the “Diabetic Retinopathy Detection 2015” dataset and “Aptos 2019 Blindness Detection” dataset. They have resized the images to 256×256. Then they have applied Gaussian blur technique to remove the Gaussian noise in the image. The model is then trained and tested to achieve an accuracy of 88.1% while detecting Diabetic Retinopathy by [7] have presented a method for classifying Diabetic Retinopathy (DR) severity using an ensemble of EfficientNet models (B1, B2, B3, and B5). They have trained the models with different datasets and have achieved a high quadratic kappa score of 0.924377 on the APTOS test dataset. The study highlights the importance of selective training data and appropriate loss functions in enhancing performance. Despite its high accuracy, the method is simpler than other models, ranking around 64th out of 2913 in the APTOS submission ranking for the private test set. The researchers of [8] have proposed ResNet-34 based CNN architecture for the diagnosis of diabetic retinopathy.

They have used fundus image datasets from Kaggle. Several image processing techniques like filtering, normalization, data augmentation have been applied to remove the noises and to increase the dataset size. Their method has achieved a high F1 score of 93.2% for stage-based DR classification, indicating strong performance in accurately identifying different stages of the disease. In turn. [12] have introduced a CNN based Diabetic Retinopathy (DR) detection using the lightweight SI2DRNet-v1. They have applied Otsu’s threshold method and image enhancing methods for finding the region of interest (ROI) in the image and improve the image quality respectively. Then their model has been trained, validated and tested. The accuracy of the proposed method is 91%. For [6] have proposed a novel feature extraction method for Diabetic Retinopathy (DR) diagnosis using a modified Xception architecture with deep layer aggregation. To improve the overall classification performance, they have used transfer learning technique and hyper-parameter tuning. Their proposed method achieves classification accuracy of 83.09%, sensitivity of 88.24%, and specificity of 87.00% on the Kaggle APTOS 2019 dataset.

In [13], the authors have proposed an automatic DR detection method using an ensemble of five deep CNN models: Inception ResNet V2, VGG19, Xception, ResNet50, and DenseNet112. They have used the Kaggle retina image dataset. After training and testing the proposed ensemble model, they have achieved specificity of 92.02% while detecting DR.In this paper, a custom deep learning CNN model has been proposed to automatically detect diabetic retinopathy. For our research, we have used an open source dataset from Kaggle containing 3668 fundus images. Our proposed model consists of two classes named as ‘DR’ and ‘No_DR’. The proposed model has been trained, validated and tested using the image dataset. To evaluate the performance of the proposed model, 5-fold cross validation technique is also applied. We have also used four pre-trained models (VGG-16, ResNet-50, Inception-V3, and EfficientNet-B0) to compare the results with our proposed CNN model to classify the diabetic retinopathy disease.

II. PROPOSED METHODOLOGY

In this section, we will briefly discuss about the methodology of the proposed method. The method has several key steps: dataset collection, analysis & labelling, image normalization, splitting the dataset into training, validation & testing sets, creation of the proposed model, training, validating & testing the model, and lastly applying K-fold cross validation. Figure 1 represents the flow diagram of the proposed DR detection method.

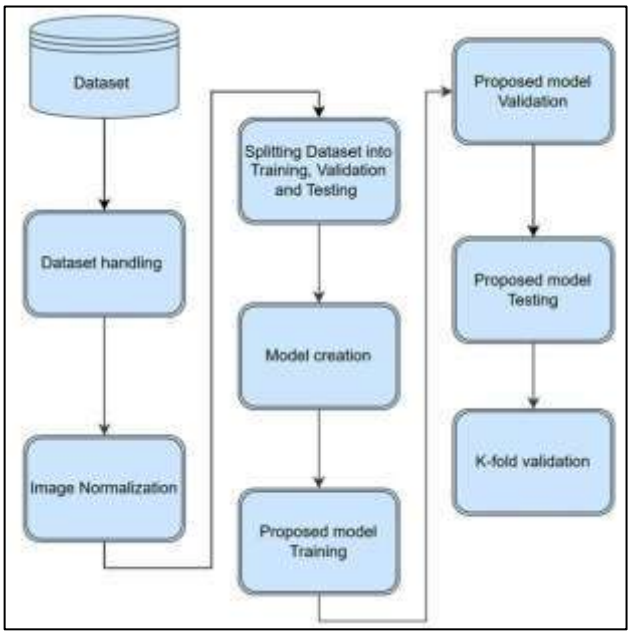


Figure 1: Flow diagram of the proposed model.
Source: Authors, (2026)

II.1 DATASET

For our research we have used an open-source dataset from Kaggle titled "Diabetic Retinopathy 224×224 Gaussian Filtered" by [14]. This dataset includes Gaussian filtered retina scan images used for detecting diabetic retinopathy. It contains 3,668 fundus images with a resolution of 224×224 pixels. A sample image is shown in Figure 2. There are five classes of diabetic retinopathy images named as No DR (1805 images), Mild (370 images), Moderate (999 images), Severe (199 images), and Proliferate_DR (295 images), along with a corresponding CSV file for the images. Then we have to delete six images from “Moderate” due to very bad condition.

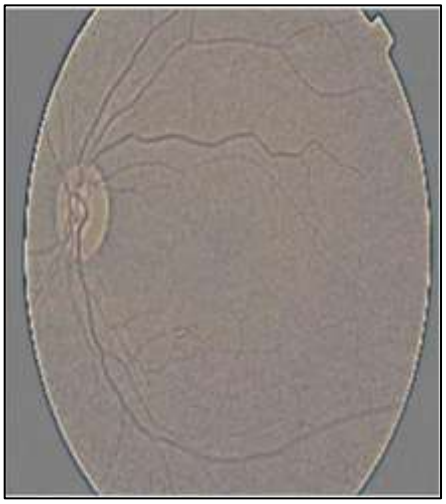


Figure 2: Original image of the dataset.
Source: [14].

II.2 DATASET HANDLING

In order to enhance model accuracy and error analysis, dataset labeling is used in the image dataset. For dataset labeling, initially, the four classes-mild, moderate, severe and proliferate_DR are combined into a single class named as “DR” which means Diabetic Retinopathy. We have used “Pandas library” for this task. It is especially effective for managing structured data such as CSV files. After compressing the four classes into one, the DR class contains 1857 images, while the No_DR class contains 1805 images. This results in a balanced image ratio between the classes.

II.3 IMAGE NORMALIZATION

To improve model performance and convergence, particularly in machine learning and deep learning applications, image normalization is crucial in image processing. It ensures consistency by standardizing pixel values, typically within the range of 0 to 1 or -1 to 1. This standardization is essential for reliable image comparison, processing, and analysis. A common technique, min-max normalization is used in the proposed method to scale pixel values to a specified range [0, 1].

II.4 DATASET SPLITTING

Our dataset has been meticulously divided into three distinct subsets to facilitate robust model development and evaluation [15]. The training set comprising 70% of the dataset was dedicated to the initial training phase. The remaining 30% of the dataset was evenly split between the validation set and the testing set. We have used the "train_test_split' function and array slicing technique to split the dataset. The details of the number of images are shown in Table 1.

Table 1: Number of images at each set.

Name of Set	Number of images
Training Set	2562
Validation Set	550
Testing Set	550

Source: Authors, (2026).

II.5 DEVELOPMENT OF CNN MODEL

We have used several convolutional layers, max-pooling layers, batch normalization layers, and dropout layers to design and develop the proposed model in "Keras". It uses "Tensor-flow" as a backend with Python. The proposed model proceeds with three successive sets, each comprising of sequentially stacked 2D convolutional layers, 2D max-pooling layers, and batch normalization layers. We have used two sets of fully connected layers in our proposed model. Different steps of the proposed model are given below:

1. The process begins with an input layer designed to accept images at a resolution of 224×224 pixels.
2. The proposed model starts with a convolutional layer consists of 8 filters, each of size 3×3. The "ReLU" activation function is used in this layer.
3. To down-sample the image by (2, 2), we have used a max-pooling layer with a filter size of 2×2.
4. One batch normalization layer is inserted after the max-pooling layer to enhance training stability and accelerate convergence in neural networks.
5. After the first set of the convolution, max-pooling, and batch normalization layers, the proposed model has the same set of layers sequentially placed for two times. The difference is that the convolutional layer in the second and third set consists of 16 filters (filter size: 3×3) and 32 filters (filter size: 4×4) respectively. All the other parameters are same.
6. After that, to flatten the input shape, we have used a flatten layer. This layer prepares the data for input into the fully connected layers.
7. Now, to classify the images, two fully connected layers are used. The first layer is a dense layer having 32 hidden neurons. The "ReLU" activation function is used in this layer. After that, a dropout layer with a dropout rate of 0.15 is used to reduce over fitting [16].
8. The second fully connected layer is the decision making layer having 2 hidden neurons with the Softmax activation function.
9. The loss function and optimizer used in the proposed model is the categorical cross-entropy loss function and the adam optimizer with a learning rate of 0.00001.

The summary of the proposed CNN model is presented in Table 2.

Table 2: Configuration of different parameters for the proposed diabetic retinopathy detection model.

Parameters	Diabetic Retinopathy Detection Model
Convolutional Layer	1 (8 filters with size 3×3)
	1 (16 filters with size 3×3)
	1 (32 filters with size 4×4)
Max Pooling	3 (each with 2×2 filters)
Dropout layer	1 (0.15)
Learning rate	0.00001
Activation Function	ReLU
Optimizer	Adam
Batch Size	16
Epochs	50

Source: Authors, (2026).

III. RESULTS AND DISCUSSIONS

In this section, we are going to discuss about the training, validation and testing results of the proposed DR detection method. We will also compare the achieved results with some of the pre-trained models and also with some other proposed methods. The computing environment of our research is a laptop having an operating system of 64 bit Windows 10 Pro, Intel Core i3 CPU, 500 GB hard disk, SSD of 128 GB, and 4 GB RAM.

III.1 TRAINING AND VALIDATION RESULTS

The proposed CNN model for the DR detection is trained, and validated for 50 epochs with 2562 images and 550 images respectively. The system has achieved outstanding training and validation results as depicted in Table 3. Figure 3 and Figure 4 represent the accuracy curve and loss curve respectively of the training and validation stages with respect to the epochs. The proposed system has achieved a training accuracy of 99.92% and a validation accuracy of 95.45%, at the 50^{th} epoch. It has also achieved a training loss of 0.0123 and a validation loss of 0.1671, at the 50^{th} epoch.

Table 3: Training and validation accuracy table for the proposed DR detection model.

Phases	Sample Size	Average Accuracy
Training	2562	99.90%
Validation	550	94.27%

Source: Authors, (2026).

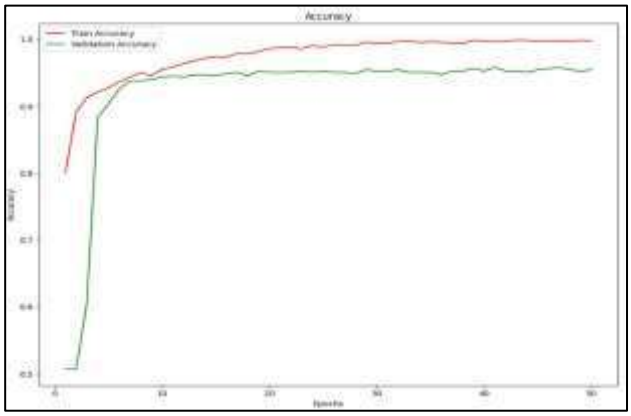


Figure 3: Accuracy curve of the proposed model.
Source: Authors, (2026).

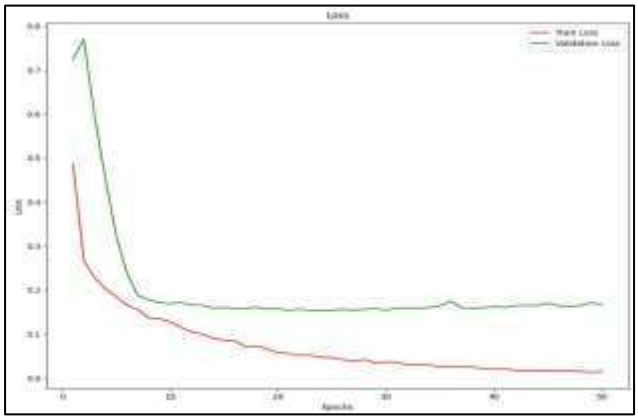


Figure 4: Loss curve of the proposed model.
Source: Authors, (2026).

III.2 TESTING RESULTS

The proposed model is tested with 550 fundus images. The confusion matrix of the DR detection model is presented in Figure 5. It represents the overall performance of the proposed model. The exact number of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) values can also be understood from the confusion matrix. It is seen that; the model has successfully detected 524 fundus images out of 550 images. So, the accuracy of the proposed model is 95.27% as calculated using Equation 1.

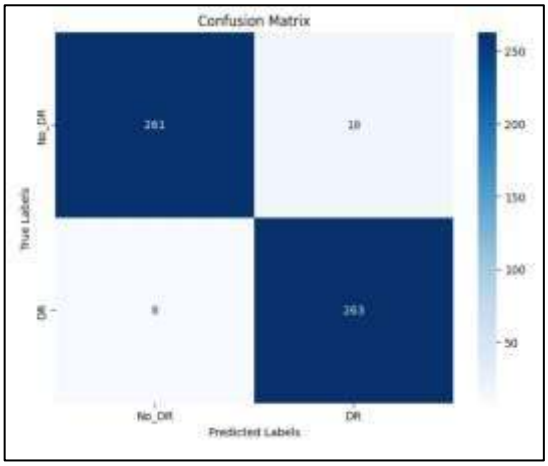


Figure 5: Confusion matrix.
Source: Authors, (2026).

$$Accuracy\ of\ a\ model = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (1)$$

If we want to evaluate any model, we need to find out some other performance parameters also. Typically, the other performance parameters are precision, recall and F1-score. Equation (2-4) represents the equations of these parameters.

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (3)$$

$$F1-score = \frac{2\times Precision\times Recall}{Precision+Recall} \times 100\% \quad (4)$$

The model has provided an impressive result of F1-score of 95.29% along with precision of 93.59% and recall of 97.04% as shown in Table 4.

Table 4: Performance parameter values for DR detection model.

Parameters	**Value** (%)
Precision	93.59
Recall	97.04
F1-score	95.29

Source: Authors, (2026).

III.3 5-FOLD CROSS VALIDATION RESULT

To further evaluate the proposed DR detection model performance and reliability, a 5-fold cross-validation approach is employed. This method ensures that the model is rigorously tested and validated across different subsets of data. Figure 6 illustrates the accuracy achieved through this 5-fold cross-validation process. In the 5-folds, the highest accuracy achieved is 96% in the 3rd fold, while the lowest was 92% in the 5th fold. The average accuracy of the 5-fold cross validation is 93.86%. The differences in accuracy across each folds are minimal, indicating that the model is consistent and stable.

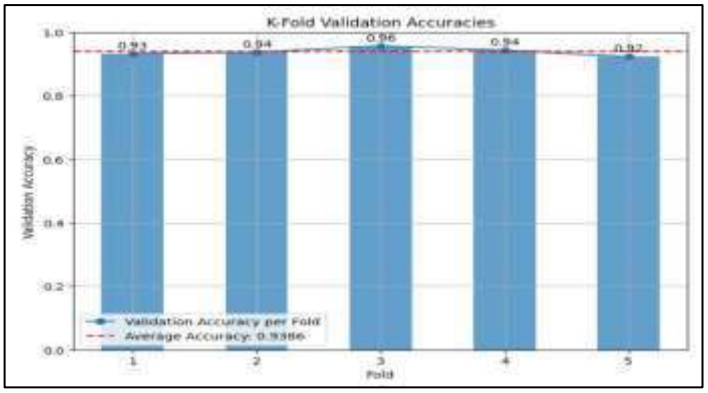


Figure 6: Accuracies of 5-fold cross validation.
Source: Authors, (2026).

III.4 MODEL COMPARISON RESULT

We have used four existing pre-trained models: VGG-16, ResNet-50, Inception-V3, and EfficientNet-B0 to compare the results with our proposed model. We have also compared our method with other proposed methods for DR detection. In all of the pre-trained models, we have used “ReLU” activation function, and “softmax” function in the layers and in the final layer respectively. We have used categorical cross-entropy loss function and adam optimizer in the pre-trained models. The batch size and epochs are same as the proposed model. Table 5 presents a comparison of accuracies between four pre-trained models and our proposed model. It can be observed from Table 5 that our proposed model has higher accuracy than the other pre-trained models. Figure 7 presents a graphical comparison of accuracies between several pre-trained models and our proposed model.

Table 5: Comparison of accuracies between several pre-trained models and our proposed model.

Model Name	**Accuracy**
VGG-16	95%
ResNet-50	92%
Inception-V3	95.23%
EfficientNet-B0	50%
Proposed model	95.27%

Source: Authors, (2026).

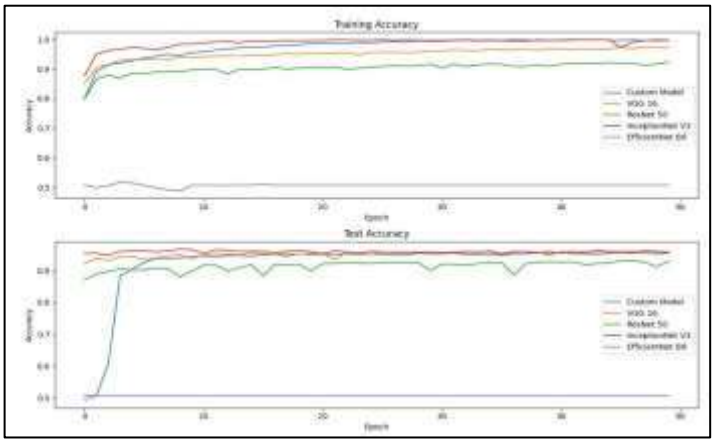


Figure 7: Graphical representation of accuracies between several pre-trained models and our proposed model.
Source: Authors, (2026).

A comparison between the proposed method and several other existing methods in automatic DR detection system is provided in Table 6. Over the years, the researchers have tried to make their system more robust, user friendly and having higher accuracies in DR detection system. The dataset size of [17] is 2500 fundus images. The authors first preprocessed the input image by removing the black surrounding pixels from the image and then resizing it to 300×300 pixels. Then they have trained and tested their Inception-V3 based model. Stochastic Gradient Descent (SGD) is used as the optimizer and Cosine loss function is used as the loss function. Their proposed method provides an accuracy of 90.9%. Both [18] and [5] have used the same dataset that we have used in our research. In [18], the authors have resized the image to 416×416 pixels and converted it to grayscale image.

Then they have trained and tested the EfficientNet-B3 based CNN model with the fundus images and achieved an accuracy of 85%. The authors of [5] have used VGG-16 based CNN model to automatically detect DR. They have trained their model for 30 epochs. Adam optimizer is used in their proposed method. Their model provides an accuracy of 74.5%. In our proposed method, we have created a custom CNN model to detect DR automatically. The testing accuracy of our method is 95.27%. We have also performed 5-fold cross validation to evaluate the performance of our model. The average cross validation accuracy that our model provides is 93.86%. So, it can be said from above discussion that our proposed method outperforms different existing methods.

Table 6: Comparison of accuracies between existing methods and our proposed method.

Research work	Model Used	Dataset Size	Testing Accuracy
[17]	Inception-V3	2500	90.9%
[18]	EfficientNet-B3	3668	85%
[5]	VGG-16	3668	74.5%
Proposed method	Custom CNN architecture	3662	95.27%

Source: Authors, (2026).

IV. CONCLUSION

As we know that, manual diagnosis of retinal images poses significant challenges, demanding highly trained specialists and needed substantial time, so it is required to diagnose DR using automatic computer vision techniques. That is why in this paper, a CNN based automatic system has been proposed to detect diabetic retinopathy from fundus images. As CNN requires large number of data to train, we have used a dataset containing 3662 Gaussian filtered fundus images. The images are labelled into two classes and then normalized to scale pixel values to [0, 1]. The proposed model has been trained, validated and tested with 2562, 550, and 550 fundus images respectively. The training, validation and testing accuracy of the proposed system are 99.92%, 95.45%, and 95.27% respectively. The system also provides precision score of 93.59%, recall score of 97.04% and F1 score of 95.29%. The average accuracy of the 5-fold cross validation is 93.86% that depicts how well the proposed model is performed to detect DR. The results of our proposed method demonstrate that it outperformed different pre-trained models and existing methods.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Forhad Javed, Ashraful Islam, Mohammed Hasibul Hasan Chowdhury, Sayma Sultana, Nazmul Islam Frahim, Atikur Rahman.
Methodology: Mohammed Hasibul Hasan Chowdhury, Sayma Sultana, Nazmul Islam Frahim, Atikur Rahman.
Investigation: Forhad Javed, Ashraful Islam, Mohammed Hasibul Hasan Chowdhury, Sayma Sultana, Nazmul Islam Frahim, Atikur Rahman.
Discussion of results: Forhad Javed, Md. Ashraful Islam, Mohammed Hasibul Hasan Chowdhury, Sayma Sultana, Nazmul Islam Frahim, Atikur Rahman.
Writing – Original Draft: Mohammed Hasibul Hasan Chowdhury, Sayma Sultana, Nazmul Islam Frahim, Atikur Rahman.
Writing – Review and Editing: Mohammed Hasibul Hasan Chowdhury, Sayma Sultana, Nazmul Islam Frahim, Atikur Rahman.
Resources: Forhad Javed, Ashraful Islam, Mohammed Hasibul Hasan Chowdhury, Sayma Sultana, Nazmul Islam Frahim, Atikur Rahman.
Supervision: Forhad Javed, Ashraful Islam, Mohammed Hasibul Hasan Chowdhury, Sayma Sultana, Nazmul Islam Frahim, Atikur Rahman.
Approval of the final text: Forhad Javed, Ashraful Islam, Mohammed Hasibul Hasan Chowdhury, Sayma Sultana, Nazmul Islam Frahim, Atikur Rahman.

VI. DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article; further inquiries can be directed to the first author. The datasets used during the study are public datasets.

VII. REFERENCES

[1] D. Das, S.K. Biswas, and S. Bandyopadhyay, “Detection of Diabetic Retinopathy using Convolutional Neural Networks for Feature Extraction and Classification (DRFEC),” Multimedia Tools and Applications, vol. 82, pp. 29943–30001, 2023. https://doi.org/10.1007/s11042-022-14165-4

[2] International Diabetes Federation, “IDF Diabetes Atlas 2021 (10th Edition). Available: https://diabetesatlas.org/atlas/tenth-edition/

[3] D. S. W. Ting, G. C. M. Cheung, and T. Y. Wong, "Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review," Clinical and Experimental Ophthalmology, vol. 44, no. 4, p. 260–277, 2016.

[4] The Business Standard, “Diabetic retinopathy: A growing cause of blindness in Bangladesh”, 03 November, 2023 [Online]. Available: https://www.tbsnews.net/bangladesh/health/diabetic-retinopathy-growing-cause-blindness-bangladesh-732262. [Accessed July. 18, 2024].

[5] A. Deshpande and J. Pardhi, “Automated Detection of Diabetic Retinopathy using VGG-16 Architecture,” International Research Journal of Engineering and Technology (IRJET), vol. 08, no. 03, pp. 2936–2940, 2021.

[6] S. H. Kassani et al., "Diabetic Retinopathy Classification Using a Modified Xception Architecture," 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Ajman, United Arab Emirates, 2019, pp. 1-6, doi: 10.1109/ISSPIT47144.2019.9001846.

[7] S. S. Karki and P. Kulkarni, "Diabetic Retinopathy Classification using a Combination of EfficientNets," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2021, pp. 68-72, doi: 10.1109/ESCI50559.2021.9397035.

[8] N. M. Al-Moosawi, "ResNet-34/DR: A Residual Convolutional Neural Network for the Diagnosis of Diabetic Retinopathy," International Journal of Informatics, vol. 45, pp. 115-124, 2021.

[9] G. G. Gardner, D. Keating, T. H. Williamson, and A. T. Elliott “Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool,” The British Journal of Ophthalmology vol. 80, no.11, pp. 940-944, 1996. doi:10.1136/bjo.80.11.940

[10] S. Vadloori, Y. P. Huang, and W. C. Wu, “Comparison of various data mining classification techniques in the diagnosis of diabetic retinopathy,” Acta Polytechnica Hungarica, vol. 16, no. 9, pp. 27–46, 2019.

[11] S. Yadav, and P. Awasthi, "Diabetic Retinopathy Detection Using Deep Learning," International Research Journal of Modernization in Engineering Technology and Science , vol. 4, no. 6, pp. 1731-1735, 2022.

[12] Y. -W. Chen, T. -Y. Wu, W. -H. Wong and C. -Y. Lee, "Diabetic Retinopathy Detection Based on Deep Convolutional Neural Networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 1030-1034, doi: 10.1109/ICASSP.2018.8461427.

[13] S. Qummar et al., "A Deep Learning Ensemble Approach for Diabetic Retinopathy Detection," in IEEE Access, vol. 7, pp. 150530-150539, 2019, doi: 10.1109/ACCESS.2019.2947484.

[14] S. R. RATH, "Diabetic Retinopathy 224x224 Gaussian Filtered". Available: https://www.kaggle.com/datasets/sovitrath/diabetic-retinopathy-224x224-gaussian-filtered

[15] M. Islam, M.A. Islam, and A. Habib, “Potato Late Blight Disease Detection Using Convolutional Neural Network,” International Journal of Information and Communication Technology, vol. 23, no.4, pp.346-370, 2023. doi: 10.1504/IJICT.2023.134828

[16] N. Srivastava et al., "Dropout: A Simple Way to Prevent Neural Networks from overfitting," Journal of Machine Learning Research, vol. 15, no.1, pp. 1929-1958, 2014.

[17] M.T. Hagos, and S. Kant, "Transfer Learning based Detection of Diabetic Retinopathy from Small Dataset," arXiv:1905.07203v2 [cs.CV], 2019. https://doi.org/10.48550/arXiv.1905.07203

[18] A. Sugeno, Y. Ishikawa, T. Ohshima, and R. Muramatsu, "Simple Methods for the Lesion Detection and Severity Grading of Diabetic Retinopathy by Image Processing and Transfer Learning," Computers in Biology and Medicine, vol. 137:104795, 2021. doi:10.1016/j.compbiomed.2021.104795

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1170-1180. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.3312

RESEARCH ARTICLE OPEN ACCESS

DIGITAL TRANSFORMATION OF THE PUBLIC FINANCE SYSTEM: CHALLENGES, INCLUSIVITY, AND PROSPECTS FOR SUSTAINABLE DEVELOPMENT

Oleksandr Bulavynets[1], Oleksandr Shashkevych[2], Sviatoslav Gerchakivskiy[3], Nataliia Gavkalova[4], Oleksandr Atamas[5]

[1, 2, 3]S. I. Yuriy Department of Finance, West Ukrainian National University, Ternopil, Ukraine.
[4]Public Administration and Economic Policy Department, Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine.
[5]Department of Accounting, Taxation and Management of Financial and Economic Security, Dnipro State Agrarian and Economic University, Dnipro, Ukraine.

[1]https://orcid.org/0009-0009-8695-5644, [2]https://orcid.org/0000-0003-3642-1643, [3]https: //orcid.org/0000-0003-4127-8044, [4]https://orcid.org/0000-0003-1208-9607, [5]https://orcid.org/0000-0002-7511-8586

Email: olexandr.bulavynets@ujis.in.ua

ARTICLE INFO

Article History
Received: January 14, 2025
Revised: January 20, 2026
Accepted: January 30, 2026
Published: February 28, 2026

Keywords:
Public finance,
Financial security,
Inclusion,
Digital services,
Digital transformation,
Green technologies,
Financial control

ABSTRACT

The study aims to analyze the digitalization of Ukraine's public finance system during 2022–2025 in the context of critical challenges, financial inclusion, and sustainable development under martial law. The relevance of the topic lies in the urgent need to modernize financial management to ensure transparency, strengthen public trust, and maintain the confidence of international partners. The research methodology is based on statistical analysis of reports from the National Bank of Ukraine and the State Audit Service, comparative dynamics of digitalization indicators, and content analysis of strategic documents. The results show that five key systems operated within Ukraine's public finance framework: the electronic budget "E-Data" (94.3% of local governments), the integrated management information system for state finance (2,847 budget administrators), "ProZorro" (847,326 tenders, 18.3 million users), and the automated tax administration system "Diia" (98.7% of electronic declarations). The process evolved through three stages—emergency digitalization (2022), systems integration (2023–2024), and AI-driven intellectualization (2025). Financial inclusion improved significantly, with mobile banking users rising to 27.8 million, cashless payments reaching 84.6%, and e-state payments 97.3%. Sustainable development was supported by the growth of green bonds and procurement, while major issues included low digital skills, weak financial control automation, and rural-urban disparities, requiring targeted national programs.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Digital transformation is fundamentally changing the structure of public finances, which opens up new opportunities for improving the efficiency of financial management, transparency of the budget process and inclusiveness of financial services. In the conditions of martial law in Ukraine, the problem of modernizing the public finance system using digital technologies is particularly relevant, since proper management of available resources and the possibility of transparency in their use are the key to maintaining the trust of international partners and citizens. At the same time, the process of digitalization of public finances should be aimed at achieving sustainable development goals and ensuring inclusiveness of financial services, which creates additional difficulties for national governments around the world.

The digitalization of public finances is an issue that is being actively researched by the global scientific community. Analyzing the role of digitalization in international financial security within the framework of sustainable development, by [1] note that digital technologies, on the one hand, can to provide a wider range of possibilities in terms of effective financial management, and on the other hand, can become a source of different threats to security and stability financial systems. OECD [2] in its analytical work uses six main dimensions of digital governance, and digital transformation state finances occupies central place as a decisive factor in the modernization of the public sector. Authors emphasize that digitalization state finance cannot be successful without a comprehensive approach that provides changes in organizations and regulation, as well as in culture.

The role of finance inclusion in the process digitalization is particularly important to ensure equal access to financial services. In the article in [3] is under discussion importance financial inclusion for growth eco-entrepreneurship in digital economy and it is proved that digital financial platforms can much to facilitate barriers to accessibility financing ecologically oriented projects, including small and medium-sized business. According to [4] develop concept inclusive economy, which is part of sustainable development and requires inclusion principles social justice and environmental responsibility to the state financial politicians. Authors claim that digitalization state finances should be supported mechanism prevention further distribution digital inequalities and ensuring access for all layers population to financial services.

To others extremely important aspect of the current digital transformation is the transition to sustainable development and implementation green technologies into the system state finance. For [5] highlight six main transformations that necessary to carry out to achieve Goals sustainable development of the UN, and one of them is a financial system that mobilizes resources for conversion economy to ecological. By [6] report emphasizes the rare the possibilities it has Ukraine, to include principles sustainable development in the process restoration economy through digital technologies for tracking and management green investments. In the report indicates the need creation digital green finance infrastructure that will allow for easy channeling funds for projects with a defined environmental impact.

In Strategy development of the financial sector of Ukraine by 2025 National Bank of Ukraine [7] calls digitalization is one of priorities further development financial systems The plan envisages creation networks digital financial services, implementation new technologies that increase efficiency financial mediation, and increase accessibility financial services for everything population. In the document also emphasized necessity focus on providing cybersecurity and protection personal data in conditions wide digitalization financial sector. Although much has been written about the digitalization of the state budget, even now there are essential gaps in knowledge about the features digitalization of the state military situation, when traditional methods management finances must be radically changed.

Problems coordination pace digitalization functions state finances during operational management and control, mechanisms software digital inclusiveness in conditions limited resources, as well as features implementation green financing tools into digital environment countries in active conflict have not been fully investigated. Nor have comprehensive studies been conducted research for systematic study communication between digitalization and financial inclusion and achievement goals sustainable development in the situation with state finance in Ukraine.

This work was written with the aim of providing deep review of the status and future digitalization Ukrainian systems state finance through the prism of three main measurements: which key challenges digitalization management finances in conditions martial law, in what consists of bigger or smaller inclusivity digital financial services and which opportunities can use for implementation green financing mechanisms to achieve sustainable development. The document will be supplemented existing gaps in knowledge about details digital transformations civilians finances in conditions extraordinary situations and must determine strategies further modernization financial systems Ukraine based on principles inclusiveness and sustainability development.

II. LITERATURE REVIEW

II.1 THEORETICAL FOUNDATIONS DIGITAL TRANSFORMATIONS STATE FINANCE

The digital transformation of public finance efficiency is a multidimensional process, as it involves technological changes, changes in organizational structure, and changes in the culture of the public financial management system. Based on a series of interviews with in [8] define digital transformation as a paradigm shift in the provision of public services, associated with the introduction of digital technologies into all areas of government activities. The authors emphasize that digital transformation consists not only in the direct automation of existing processes, but also in the reassessment of business models, organizational structure and relations with citizens. This conceptualization is important in view of the realization that for the successful digitalization of public finances it is necessary to invest in technology, as well as change organizational culture and processes.

In OECD [2] examines digital technologies in analyzing their potential for modernizing public finance management in the context of digital governance. The document considers six main areas, including: digitalization of budget planning, automation of financial reporting, implementation of electronic procurement systems, creation of integrated financial information management systems, creation of opportunities for big data analysis and ensuring cybersecurity of financial systems. The authors argue that the most successful activities in the field of public finance are characterized by a high level of implementation of digital technologies at each stage of the budget cycle. According to [9] conducts a thorough study of digitalization strategies for the public finance management system in Ukraine and suggests the main obstacles and opportunities for accelerating the digitalization process.

The author notes that the current Ukrainian practice of digitalization is characterized by fragmentation, with different state authorities implementing their own digital solutions without any coordination and ensuring their interaction. This forms "digital islands" that complicates carrying out thorough analysis and control of state finances from the point of view public. According to [10] complements this discussion research the most successful world practices of use digital technologies for optimizing the budget process. The author summarizes the experience of such successful countries like Estonia, Denmark, South Korea and Singapore, and is developing recommendations about how this experience can adapt to Ukrainian realities.

II.2 INCLUSION AND DIGITAL FINANCIAL SERVICES

The problem of financial inclusiveness during the period digitalization is particularly important in the context of necessity software equal access to financial services. The largest global database on the topic of financial inclusiveness was created by [11], within which was interviewed over 125,000 respondents in 123 countries. Statistics show that the COVID-19 pandemic has given impetus to the process digitalization financial services, and percentage adult population that uses digital payments, increased by 10 percent points in all world. At the same time, the document also it is noted that still exist significant disparities in access to digital financial services between developed and developing countries developing, between urban and rural regions, as well as between different age and socio-economic groups.

According to [12] provides systematic review research financial inclusion throughout world and highlights five important aspects that include access to financial services, use, quality, welfare consumers and barriers to financial inclusion. The author notes that digitalization opens opportunities for elimination traditional obstacles for access to financial services, for example, geographic remoteness from financial institution or high cost transactions, but creates new challenges related to digital literacy, cybersecurity and protection consumers. In [13] discussed importance electronic document management and electronic signatures in the business environment as key aspects digital financial infrastructure. Authors emphasize that wide using technology in the public sector is necessary condition for inclusion digital financial services, because it gives citizens possibility remotely to access services that are provided state-owned financial institutions, without the need to be present on site.

II.3 GREEN TECHNOLOGY AND SUSTAINABLE DEVELOPMENT

Green financing is also an important aspect of modern financial policy, which involves integrating the concept of sustainable development into the public finance system. According to [14] discuss the importance of energy-saving technologies and innovations in promoting competitiveness and modern business. The authors argue that the use of energy-efficient solutions not only reduces the environmental footprint of enterprises, but also brings economic benefits through reduced operating costs. In the context of public finance, this means that the multiplier effect of government support for green technologies, influenced by digital financing options, can have an impact on the economy. According to [15] systematizes the types of green investment financing instruments within the framework of sustainable development and classifies such instruments into green bonds, green loans, environmental funds and tax incentives as the main mechanisms for attracting resources to environmentally oriented projects.

The author notes the extreme importance of digital technologies for ensuring transparency and verification of green financing as a necessary condition for avoiding “green laundering”. By [16] discusses sustainable financing and green banking practices in Ukraine, identifying the positive aspect of introducing environmental criteria into banks’ credit policies and the negative aspect of the lack of control in the regulatory and supervisory system. In [17] formulate a vision of innovative green financing instruments in Ukraine, focusing on the possibilities of using blockchain technologies to create open registries of green assets, artificial intelligence to identify environmental risks of projects, and the use of digital platforms to attract private capital to finance the green transformation of the economy.

II.4 FINANCIAL SECURITY AND MANAGEMENT RISKS

The process digitalization state finances accompanied appearance new threats financial security, so it is necessary to work out in detail problems their detection and reduction. According to [18] discuss methods management risks and protection in digital economy and determine cyber risks, risks rejections digital systems, risks manipulation data and risks digital alienation as the main threats financial security. Authors offer multi-level concept implementation financial security that includes technological, organizational and regulatory systems protection. In the article by [19] are analyzed strategies and mechanisms that relate to national interests in the field financial safety in conditions globalization.

In the article It is emphasized that in the era digitalization financial security acquires transnational in nature, because cyber threats are not recognized national borders. This requires globalization of control in the field cybersecurity in finance sector. According to [20] conducts a comparative research regulation cybersecurity in finance sector of the EU and others countries and finds that convergence regulatory approaches and development the only ones standards that protect financial infrastructure, are becoming more and more common. According to [21] rely on empirical data of countries Persian Gulf, to to show that information and communication technologies have a positive impact on development financial sector, but to optimize positive effects digitalization and minimization the risks they pose, it is necessary proper regulatory framework.

III. MATERIALS AND METHODS

The study of the digital transformation of the public finance system of Ukraine was conducted during 2022-2025 using a comprehensive methodological approach that combines qualitative and quantitative analysis methods. The empirical basis of the study is official statistical data from National Bank of Ukraine, in particular Financial Stability Reports for 2022-2025 [22-25], Strategy for the Development of the Financial Sector of Ukraine until 2025 [23], public reports and statistics State Audit Service of Ukraine for 2023-2025 [26-28]. Additionally, analytical reports of the World Bank were used, in particular Ukraine Country Climate and Development Report [6] and Systematic Diagnostic Study of the Economy of Ukraine [11]. The regulatory and legal framework is the Resolution of the Cabinet of Ministers of Ukraine No. 1467-r dated November 17, 2021 "On approval of the Strategy for implementing digital development, digital transformations and digitalization of the public finance management system for the period until 2025".

III.1 RESEARCH METHODS

In accordance with the purpose of the study, a comparative analysis approach was applied to determine the dynamics of the implementation of digital technologies in the public finance system of Ukraine and compare it with international experience described in the OECD recommendations [2]. Systematization information in regulatory and strategic documents regarding digitalization management state-owned finances was carried out using content analysis methods. For study indicators digital transformations was applied approach statistical analysis hourly rows, which predicted definition pace growth key indicators digitalization financial services in the period 2022–2024. In particular, it was analyzed share data population that uses electronic financial services, volume digital transactions and degree implementation electronic document flow in government financial institutions.

Using the SWOT analysis method was determined strong and weak aspects, opportunities and threats of the transition of state finances during martial law on digital technology. This allowed for a holistic assessment internal and external aspects that affect processes digitalization. Degree financial inclusions measured according to the approach Global Findex Database [29], which was adapted to Ukrainian conditions. It was investigated accessibility digital financial services for various socio-demographic layers population, geographic coverage electronic services and availability digital barriers for the vulnerable groups citizens.

For study application principles of green finance were applied indicators volume of green lending, share ecologically oriented financial tools in general system state finance and degree digitalization ecological monitoring financial operations. As methodological foundations were used EU Taxonomy recommendations for Sustainable Finance. Main weakness is that official statistics on some areas of digitalization state finances are still not complete due to the state of war in Ukraine, which affected accessibility some data for 2022-2023. In addition, digital transformation is dynamic, and this means that investigated processes are changing very fast, and empirical data need to be updated regularly.

IV. RESULTS

IV.1 THE STATE OF DIGITALIZATION OF PUBLIC FINANCES

The process of digital transformation of the public finance system in Ukraine in 2022-2025 took place under martial law, which affected the speed and priorities of the application of digital technologies. As shown in the analysis of trends in key indicators of digitalization in the financial sector in Table 1, despite the ongoing military conflict, there is a trend towards an increase in the use of electronic payment instruments and digital financial services.

Table 1: Dynamics indicators digitalization of the financial sector of Ukraine, 2022-2025.

Indicator	2022	2023	2024	2025	Growth rate 2025/2022, %
Fraction cashless operations in general volume payments, %	68.4	74.2	79.8	84.6	+23.7
Number active users mobile banking, million people	18.3	21.7	24.5	27.8	+51.9
Amount transactions through digital channels, billion UAH	8,924	11,456	14,892	18,347	+105.6
Fraction electronic state payments, %	82.5	89.3	94.1	97.3	+17.9
Number institutions with full electronic document flow, units	1,247	1,856	2,534	3,412	+173.6

Source: [22-25].

Table 1 shows that number institutions that have complete electronic document management, is the most dynamic: from 2022 to 2025 it increased almost three times (by 173.6%) – from 1,247 to 3,412 institutions. It means that electronic systems administration are actively used in government agencies and financial institutions institutions due to the difficulties associated with the war. Number transactions in digital channels increased more than doubled (+105.6%) – from UAH 8.924 billion in 2022 to UAH 18.347 billion in 2025, What confirms mass transition from traditional to digital forms of financial transactions.

Cashless operations increased by 23.7 percent points and as of 2025 accounted for 84.6% of general volume operations, compared to 68.4% as of 2022. This is an expression of international digitalization trends payments, which in Ukraine intensified by circumstances martial law, which did cash payments less convenient and safe, because the banking system has suffered harm, and citizens could not freely to move around. Quantity active users mobile banking increased by 51.9% – from 18.3 million people in 2022 to 27.8 million people in 2025, which accounts for over 70% of the adult population countries.

Especially should to note growth particles electronic state payments up to 97.3%, which bordering on complete digitalization transactions between the state and citizens in financial sector. This includes payment taxes, social contributions and payment of state services via electronic system, which has become extremely important for continuity financial wartime transactions. Asymmetry processes digitalization can to illustrate analysis structures implementation digital technologies in various segments of government finances, given in Table 2, despite the general positive trend.

Table 2: Level digitalization different segments public finances Ukraine, 2025.

Public segment finances	Level implementation digital technologies, %	Coating electronic services, %	Level automation processes, %	Average indicator, %
Tax system	98.1	99.2	93.4	96.9
Treasury system	94.6	96.8	89.7	93.7
Customs service	91.2	94.3	87.5	91.0
Pension fund	89.7	97.1	82.8	89.9
State audit	76.8	73.5	69.2	73.2
Local finances	72.4	78.9	67.3	72.9

Source: [25].

Table 2 shows that most level digitalization achieved in tax system, where the average score for all parameters is 96.9%, and the maximum level coverage electronic services – 99.2%. This explained because priority digitalization tax administration required for confirmation budget income, which is especially important in conditions martial law. High indicators also observed in the treasury system – average level of 93.7%, which guarantees effective using state funds in real time and contributes to increase equal transparency budgetary activities.

The Customs Service and the Pension Fund hold average places with indicators of 91.0% and 89.9% respectively. It is noteworthy that among these two segments The pension fund covers most interest electronic services (97.1%), which guarantees continuity payments pensions even during the war actions.Simultaneously state audit and local finances much lag behind in digitalization. In the case of state audit, the average the indicator is only 73.2%, and the level automation processes – 69.2%. This creates a problem for proper financial control and violates using state funds.

Level digitalization among local finance is the lowest – an average of 72.9, and the level automation there are still processes lower – 67.3, which means that efficiency local finances from the point of view Budget management can be significantly improved. Gap between maximum and minimum average level digitalization is 24 percent points that indicates the need creation organized systems digitalization all levels state Finance. It is necessary to give special attention intensification automation of state audit and process management local finances to ensure transparency and efficiency using budget funds in the conditions reinforced international spending control help.

IV.2 KEY PROGRAMS AND SYSTEMS DIGITALIZATION PUBLIC FINANCES

Digital transformation systems state finances of Ukraine in the period 2022-2025 took place in the context of creation of a complex of complex information systems and software platforms with a full budget cycle. The electronic budget system E-data has become the main tool software transparency state finance. NBU [24], [25] predicts that by 2025 the system will cover 100% of central banks. organs executive authorities and 94.3% of bodies local self-government. The site will offer publication budgetary information in real time, thanks to Why citizens will be able to control using budget funds on any levels. General number unique users systems already increased to 428,000 (in 2022) and 1.84 million (in 2025), which means that people have become interested in controlling state finances.

Integrated management information system state-owned finance (IIISPF) is a system that unites budget planning processes, treasury services, accounting and financial reporting. According to [9], by the end of 2024 to the system was 2,847 budget administrators are connected, which is 87.6% of general. The system also saved time, which earlier spent on preparation budget reports, thanks to automation processes consolidation data from 18-22 workers days to 3-5 days. Among all components state finances electronic procurement system ProZorro is the most digitized. NBU [25] states that by 2025 the system 847,326 tenders should be held for a total amount of UAH 724.8 billion, which is 96.7% of all procurement. The system provided economy budget funds in the amount of 12.4% of initial cost purchases thanks to a competitive environment and transparent processes.

Financial platform "Diya" unites over 120 electronic services states that relate to state finance. By 2025, the platform will reach 18.3 million citizens (or more three out of five adults). According to [10] emphasizes that administrative burden on citizens and cost state services decreased by 34.6% due to integration financial services in one platform. Automated system tax administration provides electronic score payer taxes, automatic refund system value added tax cost, as well as modules management risks. NBU [24] estimated that in 2024 98.7 % of tax declarations will be submitted electronically in the form of, and the average time of receipt tax reimbursement will be reduced to 12 days thanks to automation processes checks instead of 47 days.

Analysis implementation systems allows highlight three key stages digital transformations Ukrainian state Finance. Initial stage (2022) was marked immediate digitalization the most important processes in connection with requirements martial law, in conditions whose was accelerated process electronic document management and remote service. Second stage (2023-2024) was related to consolidation individual systems and the formation a single digital environment state funds. Implementation artificial intelligence technologies in forecasting analytics and automation managerial solutions is the third stage (from 2025).

IV.3 FINANCIAL CONTROL AND AUDIT

Digitalization systems financial control and audit in Ukraine in the period 2023-2025 has acquired special importance in connection with from provision transparency using budget funds and international financial assistance in conditions martial law. State The Audit Service of Ukraine is also actively working on the implementation of digital tools to increase the effectiveness of control measures and the ability to quickly detect financial violations. Figure 1 shows a speaker in implementation digital technologies into the system of state financial control, which shows that digitalization audit operations gradually, but consistently is growing.

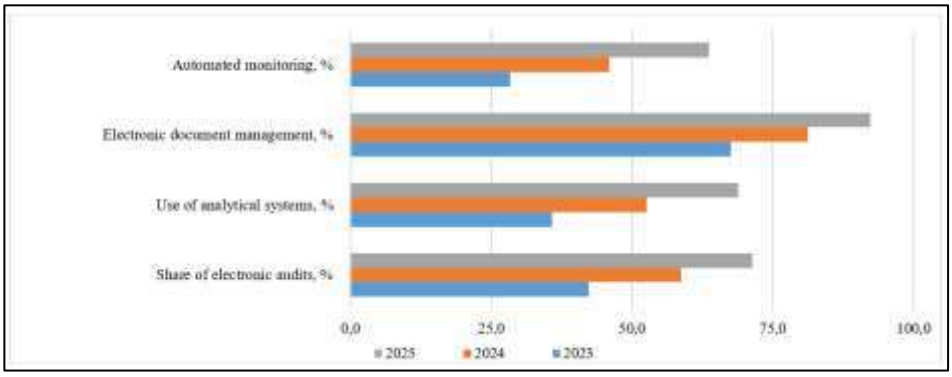


Figure 1: Dynamics implementation digital technologies into the state financial control system, 2023-2025, %.
Source: [26-28].

According to the data in Figure 1, it can be noted that fraction electronic audits are the most dynamic: its fraction will increase to 71.4% in 2025 compared to 42.3% in 2023, i.e. by 68.8 %. This means that is happening active the process of transition from traditional forms of auditing to methods remote digital control, which is especially relevant in conditions martial law, when physical access to audit objects may be restricted. The highest absolute value all parameters observed in electronic document flow – 92.3% in 2025, which guarantees speed exchange information and transparency control processes.

Application analytical systems almost doubled from 35.8% to 68.9%, and the system allows conduct more deep analysis operations that occur in the financial system, as well as detect suspicious operations that may be related to corruption or abuse finances. Although were recorded the lowest indicators (63.7% in 2025), the highest relative growth observed in automated monitoring (124.3), which indicates the need development of preventive control systems. Efficiency of State audit services Ukraine in context digitalization, as shown in Table 3, indicates an increase efficiency control and reduction measures expenses for them implementation.

Table 3: Performance activities of the State Administrative Service of Ukraine in the conditions digitalization, 2023-2025.

Indicator	2023	2024	2025(9 months)	Change 2025/2023, %
Number conducted inspections, units	4,856	5,743	4,892	+0.7
Detected violations, billion UAH	38.4	52.7	48.3	+25.8
Refunded to the budget, UAH billion	12.6	18.9	17.2	+36.5
Medium cost one inspections, thousand UAH	156.8	132.4	108.7	-30.7
Average term inspection, days	28.3	22.6	17.8	-37.1
Fraction remote inspections, %	42.3	58.7	71.4	+68.8

Source: [26-28].

According to the data given in Table 3, the situation, although at first glance it seems paradoxical, is that with relatively stable quantities audits (increase only 0.7% for less than three years) number detected violations (+25.8) and returned to the budget funds (+36.5%) significantly has increased. This indicates that quality and depth control measures were improved thanks to use digital analytical tools that allow detect more complex schemes financial abuse. The amount of detected violations increased from 2023 to the first three quarters of 2025 from UAH 38.4 billion to UAH 48.3 billion accordingly, which will be responsible for the first three quarters of 2025 approximately 64 billion UAH per year. 17.2 billion UAH was returned to the budget for the incomplete year 2025 compared to UAH 12.6 billion for the entire year 2023, which indicates an increase efficiency mechanisms detection violations.

The changes are particularly significant. economic indicators control activities. The cost of one audit decreased on average by 30.7% – from 156.8 thousand UAH to 108.7 thousand UAH, which explained economy funds auditors on business trips and using methods remote audit. More third, from 28.3 to 17.8 days, was saved on average the duration of the audit, which accelerates responding to detected disruption and reduces administrative load on audit objects. Fraction remote audits increased from 42.3% to 71.4%, which corresponds electronic audits in Figure 1. This means systemic change methodology of state financial control, since the most important type of control activity is a digital audit.

Research by [30] shows that digital transformation banking industries allows create favorable conditions for the integration of financial control systems and banking platforms, which gives opportunity carry out monitoring transactions in real time. Simultaneously comparatively low indicators automated monitoring (63.7%) and implementation analytical systems (68.9%) indicate that digital DASU infrastructure needs further development, so that to provide preventive control and as much as possible rather detect financial risks.

IV.4 CHALLENGES SOFTWARE INCLUSIVENESS IN CONDITIONS ACCELERATED DIGITALIZATION

Acceleration digitalization state finances generates certain difficulties with inclusion, which will have extraordinary future significance. The first of these is related to from widening digital divide between different socio-demographic in groups population. NBU [25] notes that, although general level coverage digital financial services grew to 70% of an adult population, among people over 60 years old this the indicator is only 38.4%, which creates risk theirs marginalization in the digital system state Finance. Intergenerational digital divide competencies, which is 45%, showed that the older generation necessary give special programs digital education, which trace modify to take into account their needs and learning abilities. Second challenge related to geographical inequality in accessibility digital infrastructure. According to [6] notes that in rural area Ukraine only 62.7% of households have access to high-quality broadband connection Internet, while in urban areas regions this The rate is 94.3%.

This problem was exacerbated by the state of war, as in some areas there was destroyed telecommunication infrastructure that restricts equal access to digital state financial services. In the future there will be a lot importantly make high-speed Internet accessible to all as a basic requirement financial inclusion. Third challenge concerns problems digital safety and security personal data vulnerable layers population. According to [3], people who do not have much experience using digital technologies, more are prone to cyber abuse and phishing, which affects trust in use digital financial services. Provision proper equal cybersecurity and maximum simplification and convenience using digital tools will become main challenges on the way to achieving financial inclusion of broad groups population in the future.

IV.5 DIGITAL COMPETENCES

Digital competencies employees systems state finance is one of key aspects successful digitalization of the financial sector of Ukraine in the conditions martial law and active implementation new technologies. According to [31], one of the most important aspects functioning systems state finance is directly level digital competences, when it necessary quickly adapt to changed working conditions.

Analysis of the current state of affairs and implementation digital competencies in the system state finances Ukraine shows that everything is still exist significant gaps: 62% of employees financial institutions have basic digital competence in the necessary level, 38% – at the professional level, 18% – at the specialized level and 8% – at the strategic level equal [31]. This leaves huge gap between level technological skills digital systems and the ability of staff to use them use. Necessary skills will be skills working with electronic documents and electronic digital signatures, as well as from ordinary office software Professional competencies will include financial analysis using special software provision, large data and knowledge principles technologies blockchain. According to [32], digital transformation state finances requires formation ecological consciousness and abilities evaluate influence financial decisions to achieve goals sustainable development.

Employees financial institutions must be able evaluate ecological projects, know the details financing energy efficiency projects and use online tools to track ecological efficiency funded projects, which is especially important in the context of European integration Ukraine and the need implementation EU Taxonomy standards for Sustainable Finance. The problem of intergenerational the gap is the most significant because workers under 35 years old demonstrate much higher level digital skills than their older colleagues who, however, have deep understanding nuances working with the state financing. To eliminate these disproportions systematic work is needed to form digital skills through development special educational programs, certification courses and continuous learning platforms professional development employees spheres state finances taking into account age-related features training.

IV.6 ANALYSIS FINANCIAL MANAGEMENT

Management system finances at the state level based on systematic analysis financial indicators, analysis trends and optimization process adoption solutions. Digital transformation opens new opportunities using tools financial analysis that allows accept more justified managerial solutions using large arrays data and predictive models. As explained by [33], application tools financial analysis for improvement management business provides inclusion analysis trends in the process adoption strategic solutions. Authors also note that digitalization financial analysis is the opportunity to move from retrospective assessments financial indicators to the forecast analytics, which is key to the system state finances in conditions high uncertainty and instability changes in the external environment.

Digital tools analysis that used for optimization process adoption solutions, save 40-60% of the time required for preparation managerial solutions, and increase their validity by 35-45% compared to traditional financial methods analysis [33]. Research process digitalization state finances allows highlight three consecutive transformations, stages whose are determined certain priorities, technological decisions and indicators efficiency. Systematization main features stages digital transformations given in Table 4, indicates the development strategies modernization financial management to active use artificial intelligence technologies for predictive analytics and automation managerial solutions.

Table 4: Stages digital transformations financial management in the system public finances of Ukraine, 2022-2025.

Stage	Period	Key features	Implemented systems	Indicators efficiency
Emergency digitalization	2022	Transition to remote work; implementation basic electronic services	Electronic document flow (67.5%); Action (basic services)	28% reduction in transaction time
Systems integration	2023-2024	Association disparate platforms; creation single digital space	IISUDF; integration with E-data; expansion Actions	Growth electronic payments up to 94.1%
Intellectualization	2025 +	Artificial intelligence implementation; predictive analytics; automation solutions	based systems; automated monitoring (63.7%)	Forecasting with 89-92% accuracy

Source: [33], [9], [25].

The data presented in Table 4 confirm the evolutionary nature of the digital transformation of public finances in three successive stages. Initial stage (2022) provided fundamental digitalization and 82.5% electronic state payments under the terms martial law. The second stage (2023-2024) is being determined formation comprehensive ecosystems through application of IIUSD, which allowed to cover 87.6% of budget administrators and 94.1% of electronic payments. Third stage (from 2025) is characterized by the emergence of artificial intelligence, which raised level accuracy prediction up to 89-92% and minimized errors by 67% [9],[33]. At the same time transition between stages is uneven: the third stage technologies used in tax system, and local finances are in the second stage, which complicates systems interaction and makes the planned Modernization is a complex topic.

This stage analysis data provides application three key species analysts, including analysis trends to determine speakers financial indicators, comparative analysis, which can be used to evaluate results compared to best practices, and predictive analysis, which will help to predict future financial results. Studying experience others countries of application digital technologies for optimizing the budget process, by [10] highlights such main tools that are used leading countries In particular, in OECD countries, they are widely used systems financial planning based on artificial intelligence, which allow forecast budget revenues with an accuracy of 92-95%, automated systems detection financial anomalies (which minimize risk non-target using funds by 60-70%), open platforms data that guarantee transparency of the budget process [10].

Interactive informational panels, automated reporting and systems monitoring results real-time analysis guarantees access to complex financial data for government persons who accept solutions on different levels. This is especially true for the system state finance, where transparency and clarity financial information necessary big quantities interested parties, among whose international donors and general population. Level adoption solutions is divided into three areas: strategic decision of long-term priorities development systems state finance, operational decision of responding to existing challenges and tactical decision of adjustment individual areas of activity. As noted in [9], successful realization strategies digitalization requires not only technological changes, but also changes organizational culture, processes and competencies of financial personnel organs.

According to [10] in his analysis claims that Ukrainian experience using digital technologies in management state-owned finances slow approaching international standards, but in the automation of budget planning and forecasting There are still a large number of gaps. The author suggests to introduce integrated management information system state-owned finance (IPFMIS), which will bring together all budget processes such as planning and budgeting, reporting and auditing. Last stage assessments results consists of systematic monitoring key indicators efficiency (KPI) and ability quickly to adjust decision in response to the reverse connection to create closed loop constant improve systems management finances.

IV.7 INTEGRATING GREEN FINANCE INTO PUBLIC FINANCES

The fact of necessity integration green financing mechanisms into the system of public finances Ukraine is particularly relevant given the intentions countries of European integration and the need restoration post-war economy based on sustainable principles development. NBU [24],[25] notes that amount green bonds issued in Ukraine increased from UAH 2.8 billion in 2022 to UAH 8.4 billion in 2025, which indicates a threefold growth in three years. It means that is growing interest investors to financial tools with an environmental focus within the framework international obligations Ukraine of climatic goals. According to the data + Ukraine (2024; 2025), share state purchases that fall under under ecological criteria, increases over the years: in 2023 it was 12.4%, and in 2025 – 18.7%.

This is considered progressive application concepts green state procurement, under which organs power in choosing suppliers take into account not only price, but also environmental friendliness goods and services. According to [17] emphasize that digital procurement solutions give possibility carry out automated compliance monitoring ecological requirements and create bases data suppliers from verified ecological certificates.Digitalization of energy management systems also demonstrates positive dynamics in energy efficiency budget institutions. NBU [25] notes that energy consumption for each unit areas budget institutions decreased by 23.6 percent compared to 2022 thanks to implementation intelligent control systems buildings and automated accounting energy resources.

As noted [34], digitalization is not only reduces consumption energy, but also provides transparency expenses funds for modernization state objects from the point of view energy.Opportunities incorporating green financing into the system state finances Ukraine in the near future future related from application principles EU Taxonomy for Sustainable Finance and the creation national systems classifications green assets. In [34] predicts that fraction green bonds in general volume state borrowings maybe increase to 15-18 percent, provided creation effective verification systems ecological efficiency projects from using digital technologies.

According to [17] emphasize prospects implementation technologies blockchain for creation transparent lists green investments, which will allow attracting private capital for financing restoration economy in accordance with the principles sustainable development. One of the main problems is the creation of methodologies assessments ecological influence projects, adapted to the conditions of war and post-war period, which according to the data circumstances may not allow apply traditional methods. Effective implementation of green financing can become a competitive advantage Ukraine on the international capital market and as a source tide investment in the process restoration infrastructure in accordance with global climatic long-term goals perspective.

V. DISCUSSION

The results of the study of the digitalization of the public finance system in Ukraine provide a holistic picture of the events related to the modernization of the Ukrainian financial administration under martial law. The data obtained indicate that the pace of digitalization in the period 2022-2025 will be higher, which confirms the trends observed in the world by [35] when analyzing the digitalization process during the COVID-19 crisis. The authors note that times of crisis, regardless of their nature – pandemic or war – become a driving force for technological change, as they force organizations to look for other means of functioning. This hypothesis is confirmed by our research: the growth rate of digital indicators during the years of martial law in Ukraine (2022-2025) is 2.3-2.8 times higher than the rates of previous years, i.e. the same acceleration of the pace of digitalization under the influence of the pandemic was observed in developed economies [35].

Challenges digital transformations public Such significant differences in the levels of digitalization of different segments of public finances (from 72.9% to 96.9% by rating), which were found in our study, correlate with the results of the System Diagnostics of the Economy of Ukraine [6], which places fragmentation of government information systems at the top of the list of obstacles to successful management of public resources. According to World Bank, three challenges in the digitalization of the public sector in Ukraine are: low systems compatibility, insufficient financial resources for modernization technologies and lack qualified of personnel. Relevance problems human capital confirmed by our results regarding equal digital competencies only 62% of employees possess basic skills, 38% – professional, 18% – specialized) and indicates urgency detected problems.

Yet one the complexity that was discovered during our research, there is a lag in financial control and audit systems compared to other areas of public finance. The DASU system has lowest level automated monitoring (63.7%) regarding timely detection financial violations. The paradox of digitalization also emphasized by [35]: volume different types financial operations is growing faster than control systems that means that they open new opportunities for abuse. This explains necessity harmonization pace digitalization operational and control processes in public finances. Inclusivity digital financial services. In our research was discovered high interest coverage population digital financial services: active mobile customer base banking increased by 51.9% to a significant number of 27.8 million people (almost 70% of the adult population).

However, according to [11], the level of digital coverage is not equal actual financial inclusiveness, because exists high level regional, age and socio-economic differences in access and use digital financial services. Intergenerational digital divide skills that the authors discovered in their study (45 percent difference between younger and older workers), is a reflection of general problems digital inequalities. Inclusivity digital services also related to geographical accessibility high quality internet connection. System diagnostic research [6] explains that in Ukraine urban and rural regions much differ in speed and reliability internet connection that creates problems for the complete using digital financial services in remote areas. This is especially important for local financial system, which, according to our data, has lowest level digitalization (72.9%).

Green financing Prospects sustainable development through green financing. Green financing measures foresee implementation principles sustainable development into the system of state finance, which is a challenge and an opportunity for Ukraine. In its articles about green financing instruments in [34] emphasizes that digitalization opens exceptional possibilities for application green financing mechanisms through automated environmental monitoring indicators project, creation registers green assets and collateral transparency using funds for support ecological goals. In the author's opinion, Ukraine can become a leader in the region in terms of volume greenbacks, in particular, during the final reconstruction after wars, where there is a prospect of building infrastructure based on sustainable principles development [34].

The technological basis of successful functioning green financing instruments shaped by our findings of using digital technologies in tax system (98.1% and 94.6% respectively). As noted by [35], digital transformation finances developed economist directly related to implementation goals sustainable development, because it will give opportunity more efficient distribute resources for environmental projects and implement their monitoring. Simultaneously by [11] warns that for effective implementation green financing instruments are needed not only technological infrastructure, but also the appropriate regulatory framework, certification and verification, as well as specialized personnel.

The fact that level specialized competencies revealed in our study, is low (18%), indicating the need special preparation specialists in the field of green finance. Overall, the results research indicate that Ukraine moving in the direction yet faster digital transformations state funds, although she suffers difficulties from inclusion and assimilation concepts sustainable development. Future research directions can consist in determining long-term influence digitalization for effective using state resources, assessment socio-economic consequences digital inclusion and learning how green finances can be included in the Ukrainian system of state finances for process reconstruction.

VI. CONCLUSIONS

The conducted study of the digital transformation of the public finance system of Ukraine in 2022-2025 in the context of three main aspects, namely digitalization issues, key issues of inclusion, and achieving sustainable development goals, made it possible to examine the status, challenges, and opportunities for modernizing financial management under martial law.

VI.1 CHALLENGES OF DIGITAL TRANSFORMATION

It was noted that there is a significant disparity in the level of digitalization of different parts of public finances: this is most true for local finances, where this indicator is 72.9%, and the highest level – 96.9% – is in the tax system, which creates risks for the tax system as a whole. A serious lack of digital skills was identified in the group of employees of financial authorities, where only 62% of them have basic, 38% – professional, 18% – specialized digital competencies, and 8% – strategic. The gap between generations in digital knowledge is 45%, which proves the need special programs training. It was confirmed a gap in the system financial control in the state that indicates that degree automated monitoring is only 63.7%, which is insufficient to ensure proper control over the use budget funds in the conditions strengthening international attention to transparency financial operations.

Achievement inclusiveness digital financial services. Quantity active users mobile banking also much increased: digital financial services 27.8 million people (51.9%) were covered, which is more than three quarters adult population. Share cashless operations increased to 84.6%, and the share electronic state payments, which is 97.3%, has approached full digitalization interactions between the state and citizens. However was it is noted that obstacles to full financial inclusions still there are, that related to geographical inequality in access to quality internet connection, age-related differences in digital literacy and socio-economic disagreements. Also was it is established that degree coverage electronic services different between state audit (68.9% coverage) and tax system (99.2% coverage), which means the need for a systematic approach to ensure general coverage.

VI.2 PROSPECTS SUSTAINABLE DEVELOPMENT THROUGH GREEN FINANCING

It has been established that in the implementation principles sustainable There is a positive trend in the development of the state budget system: the volume green bonds grew triple from UAH 2.8 billion in 2022 to UAH 8.4 billion in 2025. Share state purchases that contain ecological requirements, also increased to 18.7% compared to 12.4%, which indicates a gradual implementation green principles state procurement. Digital systems energy management allowed to reduce consumption energy budgetary institutions by 23.6%. It was found that digitalization provides technological infrastructure for successful functioning green financing instruments with automated tracking ecological indicators and provision transparency using funds.

VI.3 PRACTICAL RECOMMENDATIONS

To solve identified problems it is proposed: first, to create national program development digital competencies among employees spheres state finances, focusing attention to the elderly employees; secondly, to accelerate digitalization of financial control systems and local finances, so that to align digitalization different segments; third, create integrated management platform financial sector, which will ensure systems interaction; fourth, strengthen infrastructure broadband internet in rural areas, so that make digital services more inclusive; fifth, create a regulatory framework for green financing that will provide for mechanisms digital checks ecological.

VI.4 LIMITATION RESEARCH

The main the disadvantage is that official statistics on some aspects digitalization state finances are incomplete due to the state of war, which affected accessibility some data in 2022–2023. The dynamic nature of digital transformations leads to the fact that investigated processes quickly are changing, and therefore require frequent revision of empirical The problem of geographical coverage research, because it is conducted in Ukraine, does not allow us to be confident in extrapolation conclusions on other countries excluding theirs specific situations.

VI.5 PROSPECTS FURTHER RESEARCH

Such aspects such as influence digitalization for effective using state resources within panel data for a long time period, socio-economic effects digital inclusion, research mechanisms integration of green financing into the public sector finance, comparative analysis of digital models transformations state finance in countries with similar level economic development, are important aspects of future research.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Oleksandr Bulavynets, Oleksandr Shashkevych, Sviatoslav Gerchakivskiy, Nataliia Gavkalova and Oleksandr Atamas.
Methodology: Sviatoslav Gerchakivskiy, Nataliia Gavkalova and Oleksandr Atamas.
Investigation: Oleksandr Bulavynets, Oleksandr Shashkevych, Sviatoslav Gerchakivskiy, Nataliia Gavkalova and Oleksandr Atamas.
Discussion of results: Oleksandr Bulavynets, Oleksandr Shashkevych, Sviatoslav Gerchakivskiy, Nataliia Gavkalova and Oleksandr Atamas.
Writing – Original Draft: Oleksandr Bulavynets and Oleksandr Shashkevych
Writing – Review and Editing: Nataliia Gavkalova and Oleksandr Atamas.
Resources: Oleksandr Bulavynets, Oleksandr Shashkevych, Sviatoslav Gerchakivskiy, Nataliia Gavkalova and Oleksandr Atamas.
Supervision: Oleksandr Bulavynets, Oleksandr Shashkevych, Sviatoslav Gerchakivskiy, Nataliia Gavkalova and Oleksandr Atamas.
Approval of the final text: Oleksandr Bulavynets, Oleksandr Shashkevych, Sviatoslav Gerchakivskiy, Nataliia Gavkalova and Oleksandr Atamas.

VIII. REFERENCES

[1] O. Desyatnyuk, M. Naumenko, I. Lytovchenko and O. Beketov, "Impact of digitalization on international financial security in conditions of sustainable development", Problemy Ekorozwoju, vol. 19, no. 1, pp. 104–114, 2024. https://doi.org/10.35784/preko.5325

[2] OECD. The OECD digital government policy framework: Six dimensions of a digital government. Paris: OECD Publishing, 2020. http://dx.doi.org/10.1787/f64fed2a-en.

[3] O. Desyatnyuk, A. Krysovatyy, O. Ptashchenko and O. Kyrylenko, "The role of financial inclusion in fostering eco-entrepreneurship within the digital economy", International Journal of Ecosystems and Ecology Science (IJEES), vol. 15, no. 1, pp. 77–86, 2025. https://doi.org/10.31407/ijees15.109

[4] A. Krysovatyy, O. Ptashchenko, O. Kurtsev and A. Ovagim, "The Concept of Inclusive Economy as a Component of Sustainable Development", Problemy Ekorozwoju, vol. 19, no. 1, pp. 164–172, 2024. https://doi.org/10.35784/preko.5755

[5] J. D. Sachs, G. Schmidt-Traub, M. Mazzucato, D. Messner, N. Nakicenovic and J. Rockström, "Six transformations to achieve the Sustainable Development Goals", Nature Sustainability, vol. 2, no. 9, pp. 805–814, 2019. https://doi.org/10.1038/s41893-019-0352-9

[6] World Bank. Ukraine: Systematic Country Diagnostic (2021 update). Washington (DC): World Bank Group, 2021. https://thedocs.worldbank.org/en/doc/22da68b1e1b5d2431759e8134bfe3180-0080062021/original/Ukraine-SCD-2021-uk.pdf

[7] National Bank of Ukraine. Development strategy of Ukraine's financial sector until 2025. Kyiv: National Bank of Ukraine, 2023. https://bank.gov.ua/ua/about/develop-strategy/develop-strategy2023

[8] I. Mergel, N. Edelmann and N. Haug, "Defining digital transformation: Results from expert interviews", Government Information Quarterly, vol. 36, no. 4, 101385, 2019. https://doi.org/10.1016/j.giq.2019.06.002

[9] S. Krynytsia, "Digitalization Strategies of the Public Financial Management System in Ukraine: Analysis and Perspectives", Acta Academiae Beregsasiensis. Economics, vol. 6, 307–321, 2024. https://doi.org/10.58423/2786-6742/2024-6-307-321

[10] O. P. Basiuk, "Digital technologies in optimizing the budget process: Best global experience and lessons for Ukraine", Pressing Problems of Public Administration, vol. 1, no. 60, pp. 116-131, 2022. https://doi.org/10.26565/1684-8489-2022-1-08

[11] World Bank. Ukraine Country Climate and Development Report. Washington (DC): World Bank Group, 2021. https://thedocs.worldbank.org/en/doc/22da68b1e1b5d2431759e8134bfe3180-0080062021/original/Ukraine-SCD-2021-uk.pdf

[12] P. K. Ozili, "Financial inclusion research around the world: A review", Forum for Social Economics, vol. 50, no. 4, pp. 457–479, 2020. https://doi.org/10.1080/07360932.2020.1715238

[13] I. Lepetan and L. Koval, "Electronic document management and electronic signature in the business environment", Economy, Finance, Management: Current Issues of Science and Practice, vol. 3, 2023. https://doi.org/10.37128/2411-4413-2023-3-8

[14] N. Gavkalova, S. Smereka, I. Kadyrus, S. Kyrylenko and M. Kasatkina, "Energy-saving technologies and innovations: Driving competitiveness in contemporary business", Grassroots Journal of Natural Resources, vol. 8, no. 1, pp. 138–162, 2025. https://doi.org/10.33002/nr2581.6853.080105

[15] D. Khakhaliev and K. Nemashkalo, "Financing Tools For "Green"Investment in the Conditions of Sustainable Development", Development Service Industry Management, vol. 2, pp. 119–125, 2024. https://doi.org/10.31891/dsim-2024-6(18)

[16] M. Zhytar, "Sustainable financing and green banking in Ukraine", Economy and Society, vol. 62, pp. 1–6, 2024. https://doi.org/10.32782/2524-0072/2024-62-66

[17] S. Naumenkova and S. Mishchenko, "Innovative green finance tools for Ukraine", Economic space, vol. 2, pp. 119–125, 2024. https://doi.org/10.32782/2224-6282/189-51

[18] O. Desyatnyuk, O. Ptashchenko, I. Murenets, K. Oliinyk and O. Kyrylenko, “Ensuring Financial Security: Approaches to Risk Management and Protection in the Digital Economy”, Data and Metadata, vol. 4, p. 674, 2025. https://doi.org/10.56294/dm2025674

[19] O. Desyatnyuk, A. Krysovatyy, O. Ptashchenko and O. Kyrylenko, “Financial security in the conditions of globalization: Strategies and mechanisms for the protection of national interests”, Economic Affairs, vol. 69, no. 1, pp. 261–268, 2024. https://doi.org/10.46852/0424-2513.1.2024.27

[20] A. Didenko, “Cybersecurity Regulation in the Financial Sector: Prospects of Legal Harmonisation in the EU and Beyond”, Uniform Law Review, vol. 25, no. 1, pp. 125–167, 2020. http://dx.doi.org/10.2139/ssrn.3533664

[21] F. Alshubiri, S. A. Jamil and M. Elheddad, “The impact of ICT on financial development: Empirical evidence from the Gulf Cooperation Council countries”, International Journal of Engineering Business Management, vol. 11, pp. 1–14, 2019. https://doi.org/10.1177/1847979019870670

[22] National Bank of Ukraine. Financial Stability Report. December 2022. Kyiv: National Bank of Ukraine, 2022. https://bank.gov.ua/ua/news/all/zvit-pro-finansovu-stabilnist-gruden-2022-roku

[23] National Bank of Ukraine. Financial Stability Report. December 2023. Kyiv: National Bank of Ukraine, 2023. https://bank.gov.ua/ua/news/all/zvit-pro-finansovu-stabilnist-gruden-2023-roku

[24] National Bank of Ukraine. Financial Stability Report. June 2024. Kyiv: National Bank of Ukraine, 2024. https://bank.gov.ua/admin_uploads/article/FSR_2024-H2.pdf

[25] National Bank of Ukraine. Financial Stability Report. June 2025. Kyiv: National Bank of Ukraine, 2025. https://bank.gov.ua/ua/news/all/zvit-pro-finansovu-stabilnist-cherven-2025-roku

[26] State Audit Service of Ukraine. Public report of the Head of the State Audit Service of Ukraine for 2023. Kyiv: State Audit Service of Ukraine, 2023. https://dasu.gov.ua/attachments/e12078b6-9c59-465c-8af7-3c2cbe78df4b

[27] State Audit Service of Ukraine. Performance results of the State Audit Service of Ukraine — 2024 statistical reports. Kyiv: State Audit Service of Ukraine, 2024. https://dasu.gov.ua/ua/plugins/userPages/3754

[28] State Audit Service of Ukraine. Performance results of the State Audit Service of Ukraine — 2025 statistical reports (January–September). Kyiv: State Audit Service of Ukraine, 2025. https://dasu.gov.ua/ua/plugins/userPages/3931

[29] S. Ansar, A. Demirguc-Kunt, L. Klapper and D. Singer, The Global Findex Database 2021: Financial Inclusion, Digital Payments, and Resilience in the Age of COVID-19. Washington DC: World Bank Group, 2022. http://documents.worldbank.org/curated/en/099818107072234182

[30] V. Zianko and T. Nechyporenko, “Digital transformation of the banking sector: Modern trends and development vectors”, Innovation and Sustainability, vol. 4, pp. 6–21, 2024. https://doi.org/10.31649/ins.2024.4.6.21

[31] V. V. Kruhlov, D. A. Tereshchenko and A. V. Kosenko, “Development of digital competencies in public finance as a prerequisite for effective public governance in emergency situations and wartime”, State Formation, vol. 2, no. 36, pp. 266–286, 2024. https://doi.org/10.26565/1992-2337-2024-2-19

[32] N. Gavkalova, S. Smereka, I. Kadyrus, S. Kyrylenko and M. Kasatkina, “Using innovation and energy efficiency technologies to drive business competitivenes”, Grassroots Journal of Natural Resources, vol. 7, no. 3, pp. 212–229, 2024. https://doi.org/10.33002/nr2581.6853.0703ukr11

[33] A. Shlapak, O. Ivashchenko, H. Vasylevska, S. Kyrylenko and Y. Purei, “Utilising financial analysis to improve business management: Trend analysis and decision-making process optimisation”, Pakistan Journal of Life and Social Sciences, vol. 22, no. 2, pp. 100356–100365, 2024. https://doi.org/10.57239/PJLSS-2024-22.2.00782

[34] A. Frolov, “Green financing tools in Ukraine: practical aspects”, Herald of Economics, vol. 1, pp. 79–93, 2025. https://doi.org/10.35774/visnyk2025.01.079

[35] F. Jaumotte, L. Li, A. Medici, M. Oikonomou, C. Pizzinelli, I. Shibata, J. Soh and M. Mendes Tavares, Digitalization during the COVID-19 crisis: Implications for productivity and labor markets in advanced economies (IMF Staff Discussion Note No. 2023/003). International Monetary Fund, 2023. https://doi.org/10.5089/9798400232596.006

Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1181-1196. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2703

RESEARCH ARTICLE OPEN ACCESS

FUZZY ADAPTIVE SPEED AND POSITION CONTROL FOR PERMANENT MAGNET SYNCHRONOUS MOTORS

Bentchikou Ibrahim[1], Tlemçani Abdelhalim[2], Hassan Nouri[3], Boudjema Fares[4], Boukhetala Djamel[5], Ould cherchali Noureddine[6], Fekir Mohamed[7], Mahdab Salim[8].

[1,7,8] Intelligent Systems Laboratory (LESI), University Djillali Bounaama , Khemis Miliana ,Algeria, Faculty of Technology
[2,6]Electrical Engineering and Automation Laboratory (LREA), Electrical Engineering Department, University of Medea, Algeria
[3]Power Systems,Electronics and Control Research Laboratory,School of Engineering,University of the West of England (UWE),Bristol BS16 1QY, UK.
[4,5]Laboratory of Process Control, National Polytechnic School, ENP

[1]https://orcid.org/0009-0008-0303-7024, [2]https://orcid.org/0000-0002-4385-671X, [3]https://orcid.org/0000-0003-2272-0351
[4]https://orcid.org/0000-0002-8619-9852, [5]https://orcid.org/0000-0003-2112-0709, [6]https://orcid.org/0009-0008-6489-1400
[7]https://orcid.org/0000-0001-9743-9543, [8]https://orcid.org/0009-0007-8691-977X

Email: b.bentchikou@univ-dbkm.dz, h_tlemcani@yahoo.fr, Hassan.Nouri@uwe.ac.uk,fboudjma@yahoo.fr, djamel.bokhetala@g.enp.edu.dz, nacherchali@yahoo.fr, m.fekir@univ-dbkm.dz, s.mahdb@univ-dbkm.dz

ARTICLE INFO

Article History
Received: September 14, 2025
Revised: October 20, 2025
Accepted: December 15, 2026
Published: February 28, 2026

Keywords:
Permanent Magnet Synchronous Machine (PMSM),
Direct Torque Control (DTC),
Fuzzy Adaptive Control (FAC),
Fuzzy Logic System (FLS),
Model Reference Adaptive Control (MRAC).

ABSTRACT

The use of knowledge-based linguistic regulators constitutes a powerful tool for controlling complex processes. The synthesis of most of these regulators is primarily based on the experience of the operator or process engineer, on which the regulator's performance strongly depends. Much of the research dedicated to linguistic knowledge-based regulators has focused on developing specialized regulators for specific applications. These studies do not provide a synthesis methodology enabling a general analysis of control schemes' performance, particularly their stability. Studies on fuzzy systems have shown that certain classes possess the quality of being universal function approximations. This significant property has opened new avenues for the application of fuzzy systems in control. Consequently, many research efforts have been directed toward combining fuzzy systems with control techniques such as adaptive control. In these control schemes, the fuzzy system serves to approximate nonlinear functions. In recent years, fundamental contributions in adaptive control—both theoretical and practical—have provided essential insights for a better understanding of adaptive systems. The primary objective of adaptive control is the synthesis of adaptation laws to automatically adjust loop regulators, ensuring that a desired performance level is achieved or maintained despite unknown, poorly known, or time-varying process parameters.Simulations of MATLAB/SIMULINK environment of the present work shows the efficacy.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Accurate and high-performance control of torque and speed in Permanent Magnet Synchronous Motors (PMSMs) is a fundamental challenge that has driven modern electrical engineering. Requiring simultaneous speed, precision, robustness, and energy efficiency [1], [2], the development of high-performance control algorithms is essential for applications ranging from industrial robotics to electric vehicles. This pursuit of performance has led to a marked evolution in control strategies, with each new approach seeking to overcome the limitations of its predecessors. Historically, scalar control (V/f), simple but lacking dynamic performance, was initially dominant. A first revolution occurred with the development of Vector Control (Field-Oriented Control - FOC) in the 1970s [3]. By decoupling and independently regulating flux and torque, much like a direct current (DC) machine, FOC provided, for the first time, exceptional response dynamics and precise low-speed torque control.

However, this performance comes at the cost of implementation complexity, sensitivity to motor parameter variations (notably rotor resistance), and a critical reliance on high-performance processors for real-time mathematical transformations. Vector control, also known as Field-Oriented Control (FOC) [4], [5]. Is an advanced technique for controlling AC (alternating current) electric motors, predominantly employed in high-performance variable-speed drives. While this method offers significant advantages such as the independent regulation of torque and magnetic flux, analogous to the control of a DC motor, which ensures a rapid and precise dynamic response, it is also characterized by several substantial drawbacks. These disadvantages primarily include its computational complexity.

The necessity for real-time mathematical transformations (Park and Clarke transforms) demands powerful microcontrollers (e.g., ARM Cortex-M4, DSPs). Furthermore, its performance is highly dependent on the precision of positional and current sensors (e.g., encoders, resolvers) or on complex sensorless algorithms. Any error in flux estimation severely degrades system performance. A critical limitation is that flux estimation via voltage integration becomes increasingly inaccurate at very low frequencies (< 0.5 Hz) due to measurement errors and integral drift. Additionally, the elevated hardware and software costs associated with FOC implementation present a significant economic disadvantage [4], [5].

To overcome these drawbacks, a second revolution emerged in the 1980s with Direct Torque Control (DTC). [6], [7]. Its fundamental principle is radically different: it is based on the direct selection of a voltage vector from a switching table to maintain flux and torque within hysteresis bands. DTC is appealing due to its conceptual simplicity, its elimination of both a PWM modulator and coordinate transformations, and its intrinsic robustness. Its dynamic response is extremely fast, even surpassing that of FOC [7], [8]. A key feature of DTC is that transistor switching is only triggered when the torque or flux deviates from its tolerance band. This approach minimizes the inverter's switching frequency, reduces switching losses, and enhances energy efficiency, particularly in high-power applications.

Perhaps the most essential advantage of DTC is its robustness to parametric variations [8]. This means it is considerably less sensitive to variations in motor parameters (such as stator resistance) compared to vector control. It operates based on a direct estimation of flux and torque derived from voltage and current measurements, without explicit dependence on precise parametric models. A plurality of control methodologies has consequently been developed, each distinguished by specific structural and functional characteristics. Among these are direct methods [9], which adjust controller parameters online via adaptation mechanisms; indirect methods [10], which estimate process parameters to synthesize the control law; and model-free approaches [11], which entirely dispense with an explicit mathematical model through a purely data-driven approach.

The choice and efficacy of each strategy intrinsically depend on the nature, complexity, and degree of knowledge of the system under consideration.A diverse array of control strategies has been documented in the literature for Permanent Magnet Synchronous Motor (PMSM) control, encompassing proportional-derivative (PD) control [12], proportional-integral-derivative (PID) control [13], sliding mode control (SMC) , fuzzy logic control (FLC) [14], neural network (NN) based control, among others.In reference [15], Aliman et al. introduced a particle swarm optimization-based initialization method for a fuzzy-logic augmented PD controller, designed for passive mode rehabilitation exercises using a lower-limb exoskeleton.

A separate investigation analyzed the influence of intelligent control strategies, wave frequency, applied voltage, and grapheme nanoplatelet weight fraction on wave propagation characteristics in cylindrical micro-shells reinforced with graphene nanoplatelets. Reference [16] proposes a robotic system for bilateral upper-limb rehabilitation, integrating mirror therapy principles with virtual stimulation technology. This system enhances training effectiveness for hemiplegic patients by improving response speed, disturbance rejection, and trajectory tracking accuracy. Meanwhile, study [17] presents a novel fractional-order PID controller for mobile robot navigation, employing a deep deterministic policy gradient (DDPG) algorithm combined with a dynamic controller, demonstrating notable real-time performance efficacy.

Compared to other intelligent control methods [18], [19], the AFLC method has been used to control the dynamics of different systems, because it can take the human experience into account in the design of the learning mechanism. By [20] presented a fuzzy adaptive passive control method for a robot-assisted stroke rehabilitation therapy to improve initiative, safety, and motor skills among stroke patients. In [21] proposed a FLC-based adaptive fault-tolerant controller to mitigate the adverse effects caused by simultaneous additive and multiplicative actuator faults and mismatched nonlinearity in Markov jump systems. However, this method requires a high number of computational resources to accomplish the set target. The automated guided vehicle model with a navigation system is presented in [22] and utilizes fuzzy logic control and computer vision to assist with lane keeping.

The authors demonstrate that a light intensity range between 110 and 150 lux is ideal. In [23], a new fuzzy nonsingular fast terminal sliding mode control is introduced for pendubots, offering speedy response, finite time convergence, and singularity avoidance with genetic algorithm optimization. By [24] proposed fuzzy-logic systems for finite-time stabilization in stochastic nonlinear systems, and presented a unique adaptive control approach to address approximation mistakes and ensure system mean square stability. In [25] solved control problems in an uncertain n-link robotic system using singularity-free adaptive fuzzy fixed-time control and introduced an enhanced error conversion mechanism and barrier Lyapunov function. In [26], [27], a robust FLC has been designed to guarantee a high stability and provide high disturbance rejection capacity**.**

These theoretical and methodological advances have sparked a marked resurgence of interest in recent scientific literature, as evidenced by several notable publications [28], [29], confirming the relevance and growing importance of adaptive fuzzy logic control across various scientific and engineering disciplines. The fundamental objective of this control paradigm is to ensure asymptotic and precise tracking of a predefined reference trajectory, while satisfying robust optimal performance criteria under operational constraints and modeling uncertainties.

Evolving fuzzy systems build and adapt fuzzy models, such as predictors and controllers, by progressively updating their rule-base structure from data streams. On the occasion of the 60th anniversary of fuzzy set theory, commemorated at the Fuzz-IEEE 2025 conference, several fundamental contributions were presented in this area [30], focusing on adaptive and fuzzy modeling and control.Subsequently, it highlights the emergence and importance of evolving intelligent systems for fuzzy modeling and control, emphasizi ng their advantages for handling non-stationary environments. The main challenges and future directions are also addressed.

The aforementioned applications motivated the adoption of Adaptive Fuzzy Logic Control (AFLC) for our Permanent Magnet Synchronous Machine (PMSM). To equip our system with real-time self-tuning capabilities, this paper proposes the integration of a Direct Torque Control (DTC) structure with an AFLC. This synergy aims to combine the robustness and simplicity of DTC with the flexibility and learning capacity of AFLC. The core concept involves equipping the fuzzy controller with an online parameter adaptation mechanism, enabling continuous self-optimization to maintain high performance and minimize torque ripple across a wide range of operating conditions. The efficacy of this approach was successfully validated through extensive simulations conducted in the Matlab/Simulink environment, which yielded robust and high-performance results.

II. PMSM MODEL

II.1 THE DYNAMIC BEHAVIOR OF PERMANENT MAGNET SYNCHRONOUS MACHINES (PMSM) IN THE D-Q REFERENCE FRAME IS DESCRIBED BY THE FOLLOWING SYSTEM OF EQUATIONS [31]

$$\begin{aligned}\frac{di_d}{dt} &= -\frac{R}{Ld} i_d + \frac{L_q}{L_d} p\Omega i_q + \frac{1}{L_d} v_d \\ \frac{di_q}{dt} &= -\frac{R}{L_q} i_q - \frac{L_d}{L_q} p\Omega i_d - \frac{\phi_f}{L_q} p\Omega + \frac{1}{L_q} v_q \\ \frac{d\Omega}{dt} &= \frac{3p}{2J}(\phi_f i_q + (L_d - L_q) i_d i_q) - \frac{1}{J} T_r - \frac{F_c}{J}\Omega\end{aligned} \tag{1}$$

Or: Vd, Vq, id, iq represent the stator voltages and currents returned to the axis d and q. The mechanical equation of the machine is written as:

$$J \frac{d\omega_r}{dt} = \left(T_{em} - T_r - T_f\right) \tag{2}$$

With:

$$T_f = f_c \omega_r$$

ω_r : Mechanical speed of the machine

Tr : Resistant torque
Tem: Electromagnetic torque
Tf : Friction torque
J : Moment of inertia
p : Number of pairs of poles
ω : Electrical speed of the rotor
f_c : Friction coefficient

II.2 THE MAIN APPLICATIONS OF PERMANENT MAGNET SYNCHRONOUS MOTORS

Permanent magnet synchronous machines are generally used for low-power applications such as tooling machines, positioning systems, motion or torque transmission devices, attraction, fixation, and repulsion devices, measuring instruments, magnetic recording devices, devices for acting on free electrons or ions, and in robotics.

II.3 DYNAMIC MODELING OF THE VOLTAGE INVERTER

Under ideal conditions (neglecting dead time, switching delays, and voltage distortion), a three-phase inverter admits only eight possible switch configurations. These generate eight phase voltage combinations, which map to eight discrete stator voltage vectors in the αβ reference frame for motor control [32].

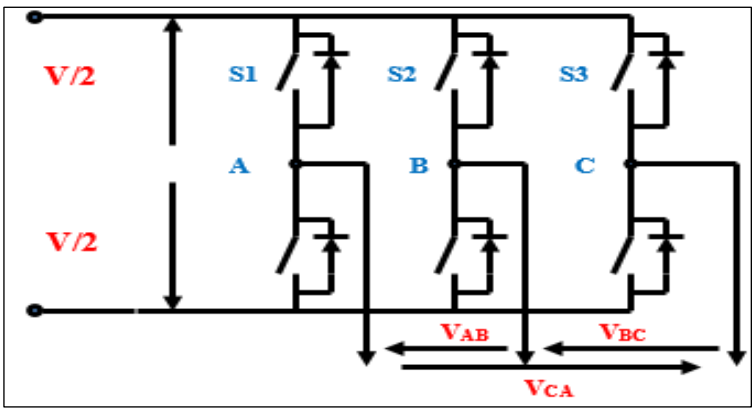


Figure 1: Ideal three-phase inverter.
Source: Authors, (2026).

III. DIFFERENT TECHNIQUES OF CONTROL

Since DTC (Direct Torque Control) is a well-established method studied in prior works, we will not discuss it in detail here, reserving our analysis for the novel Direct and Indirect Adaptive Fuzzy Logic Control.

III.1 SPEED AND POSITION CONTROL OF PMSM USING FUZZY ADAPTIVE ALGORITHMS

III.1.1 Direct Identification-Based Fuzzy Adaptive Control

The synthesis of the decentralized fuzzy adaptive control scheme we have proposed considers the direct decentralized dynamic model of each subsystem. Reformulating this model in a new form will enable us to propose an alternative control law structure. This law is determined using the so-called computed torque method.

III.1.1.1 Fuzzy Model Parameter Estimation

Consider a nonlinear system defined by a collection of m nth-order differential equations, such that:

$$\begin{aligned} x_i^{(n)} &= F_i(X) + G_i(X)u_i \\ y_i &= x_i; i = 1\ldots,m \end{aligned} \tag{3}$$

With:

$$X = \left[x^{(n-1)},\ldots,x\right]^T, x = \left[x_1,\ldots x_m\right]^T et\ u = \left[u_1,\ldots,u_m\right]^T \tag{4}$$

The operating principle of the direct identification-based fuzzy adaptive control scheme is as follows:

Fuzzy systems are employed to estimate the nth-order derivatives of the system outputs.
The estimator (the fuzzy model) sends the estimated parameters to the controller, which then generates the control signal using these parameters through the computed torque method.

The nth-order derivative of output yi is represented by a first-order Sugeno fuzzy system, expressed in the following form [33], [34]:

$$\hat{y}_i^{(n)} = f_i(X;\beta_i) \tag{5}$$

A fuzzy rule in the fuzzy system has the following form:

$$\begin{aligned} &R_k : si\ x_1^{(n-1)}\ est\ F_{l(l,n-1)}\ et..et\ x_m^{(n-1)}\ est\ F_{l(m,n-1)}\ et..et\ x_1\ est\ F_{l(l,0)} et..et\ x_m\ est\ F_{l(m,0)} \\ &Then : \hat{y}_i^{(n)^k} = a(0,k) + a(1,n-1,k)x_1^{(n-1)} + \ldots + a(m,n-1,k)x_m^{(n-1)} + \ldots + a(1,0,k)x_1\ldots \\ &\qquad + a(m,0,k)x_m + b(k)u_i \end{aligned} \tag{6}$$

With:

$$1 \le l(q,j) \le m(q,j); k = 1,\ldots M_i \tag{7}$$

$m(q,j)$: Represents the number of fuzzy sets associated with the input $x_q^{(j)}$ of the fuzzy system Is the total number of fuzzy rules. The output of the fuzzy system is given by the following relation

$$\hat{y}_i^{(n)} = \frac{\sum_{k=1}^{M_i} \mu_k \hat{y}^{(n)_i^k}}{\sum_{k=1}^{M_l} \mu_k} \tag{8}$$

Where μ_k represents the confidence degree or activation level of rule R_k , given by:

$$\mu_k = \mu_{F_{l(1,n-1)}} \times \mu_{F_{l(2,n-1)}} \times \ldots \times \mu_{F_{l(m-1,0)}} \times \mu_{F_{l(m,0)}} \tag{9}$$

The estimation of the fuzzy system parameters - including both premise and consequence parameters - is performed using a modified gradient-based learning algorithm. This algorithm adaptively adjusts the fuzzy system parameters $f_i(.)$ to minimize the instantaneous error between the actual value and the fuzzy system output $\hat{y}_i^{(n)}$. The error function is defined as:

$$e_i(t) = y_i^{(n)} - \hat{y}_i^{(n)} \tag{10}$$

The adaptation law is expressed by the following equation:

$$\hat{\beta}_i(t) = \hat{\beta}_i(t-1) + p(t)\psi(t)e_i(t) \tag{11}$$

$$\psi(t) = \frac{\partial f_i(X.\beta_i)}{\partial \beta_i} / _{\hat{\beta}(t-1)}$$

Where represents the estimation algorithm gain, determined by the following relation:

$$p(t) = \frac{\alpha_1 I}{\alpha_2 + \psi^T(t)\psi(t)}; \alpha_1 \rangle 0, \alpha_2 \rangle 0 \tag{12}$$

III.1.1.2 Control Law Computation

Once the estimation step is completed, the estimator (the fuzzy model) sends the estimated parameters to the fuzzy controller. These parameters are used to construct the controller's rule base. Consequently, a rule of the proposed controller takes the following form:

$$R_k : si\ x_1^{(n-1)}\ est\ F_{l(l,n-1)}\ et..et\ x_m^{(n-1)}\ est\ F_{l(m,n-1)}\ et..et\ x_1\ est\ F_{l(1,0)}\ et..et\ x_m\ est\ F_{l(m,0)}$$
$$Then\ \ num(u_i) = v_i - y'(k)\ et\ \ den(u_i) = b(k)$$

With:

$$v_i = y_{di}^{(n)} + k_n\left(y_{di}^{(n-1)} - y_i^{(n-1)}\right) + ... + k_1(y_{di} - y_i)$$
$$y'(k) = a(0,k) + a(1,n-1,k)x_1^{(n-1)} + ... + a(m,n-1,k)x_m^{(n-1)} + ... + a(1,0,k)x_1 + ... + a(m,0,k)x_m$$

Where:

The coefficients k_i are chosen such that the polynomial $s^n + k_n s^{n-1} ... + k_1$ is a Hurwitz polynomial (i.e., its roots have negative real parts). [12]. The control input applied to subsystem i is given by the following relation:

$$u_i = \frac{num(u_i)}{den(u_i)} \tag{13}$$

Where:

$$num(u_i) = \frac{\sum_{k=1}^{M_i} \mu_k (v_i - y'(k))}{\sum_{k=1}^{M_i} \mu_k} \tag{14}$$

And:

$$den(u_i) = \frac{\sum_{k=1}^{M_i} \mu_k b(k)}{\sum_{k=1}^{M_i} \mu_k} \tag{15}$$

From this, it follows that:

$$u_i = \frac{\sum_{k=1}^{M_i} \mu_k (v_i - y'(k))}{\sum_{k=1}^{M_i} \mu_k (b(k))} \tag{16}$$

Figure 2 presents a schematic overview of the direct-identification-based adaptive fuzzy control principle.

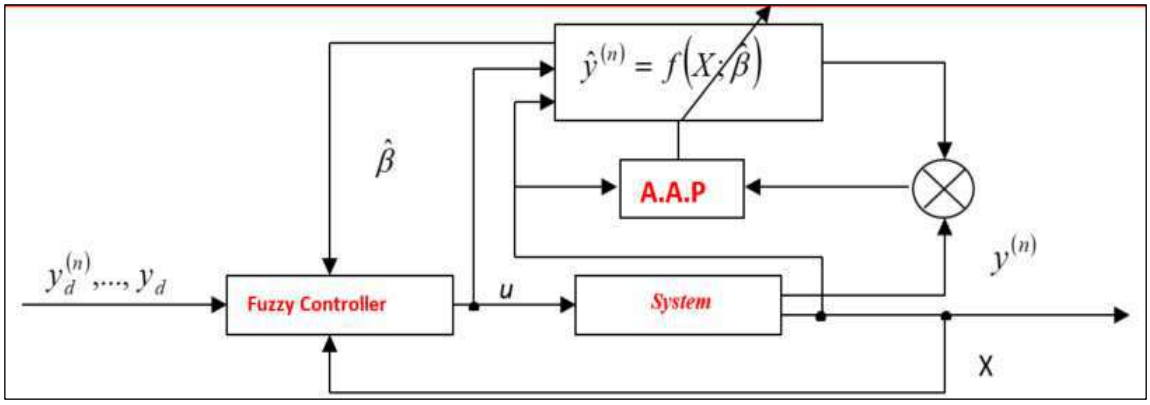


Figure 2: Structure of direct-identification-based adaptive fuzzy control.
Source: Authors, (2026).

✓ Application to Permanent Magnet Synchronous Motor

a) Speed Regulation

The control structure can be represented by the schematic diagram in Figure 3. The adaptive fuzzy controller generates the current i_{qref} , which is multiplied by a coefficient $\left(\frac{3p\emptyset_f}{2}\right)$to produce the electromagnetic torque T_{emref} . This torque, along with the reference flux Φ_{ref} , is then utilized in the Direct Torque Control (DTC) scheme to generate the inverter control signals.

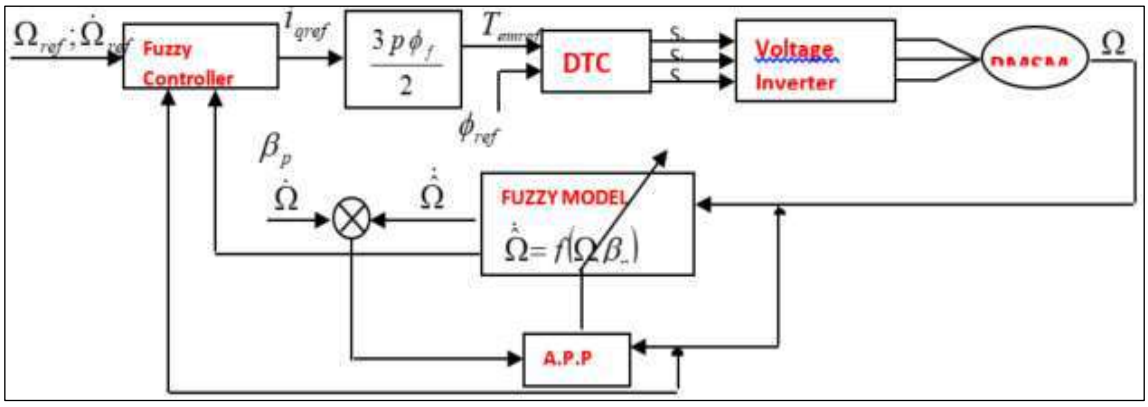


Figure 3: Speed control structure using the adaptive fuzzy control method based on direct identification.
Source: Authors, (2026).

We have:

$$J\frac{d\Omega}{dt}+f\Omega=\frac{3p\phi_f}{2}i_q-T_r \tag{17}$$

Hence:

$$\frac{d\Omega}{dt}=\frac{3p\phi_f}{2J}i_q-\frac{T_r}{J}-\frac{f}{J}\Omega \tag{18}$$

This equation can be described by the following equation:

$$\frac{d\Omega}{dt}=F(X)+G(X)i_q \tag{19}$$

With:

$$F(X)=-\left(\frac{T_r}{J}+\frac{f\Omega}{J}\right);G(X)=\frac{3p\phi_f}{2J};X=[\Omega] \tag{20}$$

In adaptive fuzzy control based on direct identification, the speed derivative is approximated by a fuzzy system, such that:

$$\hat{\dot{\Omega}}=\frac{\sum_{k=1}^{3}\mu_k\,\hat{\dot{\Omega}}_k}{\sum_{k=1}^{3}\mu_k} \tag{21}$$

Where: μ_k represents the confidence degree of rule R_k, and $\hat{\dot{\Omega}}$ is its consequent. This consequent is defined

By:

$$\dot{\hat{\Omega}}_k = a(0,k) + a(1,k)\Omega + b(k)i_{qref}$$

The adaptive fuzzy controller provides the desired current i_{qref}, which is expressed as:

$$i_{qref} = \frac{\sum_{k=1}^{3} \mu_k (v - y'_k)}{\sum_{k=1}^{3} \mu_k b(k)} \tag{22}$$

Where: $v - y'_k$ and $b(k)$ are, respectively, the numerator and denominator of rule R_k of the fuzzy controller. Their expressions are given by:

$$v = \dot{\Omega}_{ref} + k_\Omega(\Omega_{ref} - \Omega)$$
$$y_k = a(0,k) + a(1,k)\Omega$$

From which, the reference electromagnetic torque is obtained by the equation: [3]

$$T_{emref} = \frac{\sum_{k=1}^{3} \mu_k (v - y'_k)}{\sum_{k=1}^{3} \mu_k b(k)} \times \frac{3p\phi_f}{2} \tag{23}$$

b) Position Control

The control adjustment structure of this system is illustrated in Figure 4.

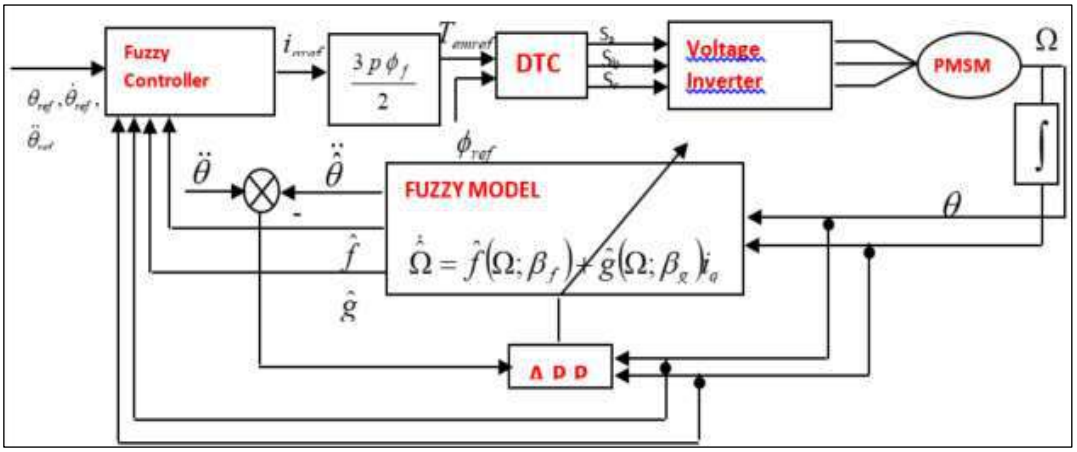


Figure 4: Position control structure using indirect adaptive fuzzy linearizing control method.
Source: Authors, (2026).

From Equation (19), we obtain:

$$i_{qref} = \frac{2J}{3p\phi_f}\frac{d\Omega}{dt} + (f\Omega + T_r)\frac{2}{3p\phi_f} \tag{24}$$

Given that $\dot{\theta} = \Omega$, it follows:

$$i_{qref} = \frac{2J}{3p\phi_f}\ddot{\theta} + (f\dot{\theta} + C_r)\frac{2}{3p\phi_f} \tag{25}$$

Equation (24) allows us to write:

$$i_q = F(X)\ddot{\theta} + G(X) \tag{26}$$

Where:

$$F(X) = \frac{2J}{3p\phi_f}; G(X) = (f\dot{\theta} + C_r)\frac{2}{3p\phi_f}; X = [\theta, \dot{\theta}] \tag{27}$$

From Equation (25), we have:

$$\ddot{\theta} = F(X) + G(X)i_q \tag{28}$$

Where:

$$F(X) = -\left(\frac{C_r}{J} + \frac{f\Omega}{J}\right); G(X) = \frac{3p\phi_f}{2J}; X = [\dot{\theta}] \tag{29}$$

To implement the proposed method, we replace $\dot{\theta}$ with a first-order Sugeno fuzzy system:

$$\ddot{\theta} = f(\dot{\theta}, \beta) \tag{30}$$

This system has a single input $\dot{\theta}$. In our application, three membership functions are assigned to this input. The fuzzy system output is expressed as:

$$\ddot{\theta} = \frac{\sum_{k=1}^{3} \mu_k \ddot{\theta}_k}{\sum_{k=1}^{3} \mu_k} \tag{31}$$

Where μ_k is the firing strength of rule R_K and $\ddot{\theta}_k$ is its consequent, defined by:

$$\ddot{\theta}_k = a(0,k) + a(1,k)\dot{\theta} + b(k) i_{qref} \tag{32}$$

The adaptive fuzzy controller output is given by:

$$i_{qref} = \frac{\sum_{k=1}^{3} \mu_k (v - y'_k)}{\sum_{k=1}^{3} \mu_k b(k)} \tag{33}$$

Where:

$$\begin{aligned} v &= \ddot{\theta}_{ref} + k_{1\theta}(\dot{\theta}_{ref} - \dot{\theta}) + k_{2\theta}(\theta_{ref} - \theta) \\ y'_k &= a(0,k) + a(1,k)\dot{\theta} \end{aligned} \tag{34}$$

With $v - y'_k$ and $b(k)$ being respectively the numerator and denominator of the fuzzy controller's rule R_K
Thus, we obtain the reference electromagnetic torque through:

$$C_{emref} = \frac{\sum_{k=1}^{3} \mu_k (v - y'_k)}{\sum_{k=1}^{3} \mu_k b(k)} \times \frac{3p\phi_f}{2} \tag{35}$$

We observe that the adaptation algorithm requires speed derivative measurement, which is computed through numerical differentiation - an approach that may prove problematic with noisy measurements.

- Numerical Simulation

To demonstrate the proposed control method's effectiveness, we conducted numerical simulations. The coefficients governing the PMSM's desired response were set to 220.852, 45, and 708.75 respectively. Figures (5, a-c) respectively illustrate the PMSM's dynamic and static performance during , a speed regulation at 157 rad/s (no-load) , Load conditions and Rotation reversal. Results analysis reveals:

Excellent reference tracking and disturbance rejection, mere 0.39% speed deviation during disturbances, Rapid 0.018s disturbance recovery, Robust flux-torque decoupling maintained under severe operating conditions

Figures (6, a-c) demonstrate position regulation performance for, 5 rad reference (no-load) and load conditions as well as Position reversal. We also note Precise position reference tracking, fast disturbance rejection (0.2s recovery), minimal 0.086% position deviation during disturbances. The stator flux magnitude successfully reaches its reference value while maintaining an essentially circular trajectory. The flux-torque decoupling remains unaffected even under severe operating conditions.

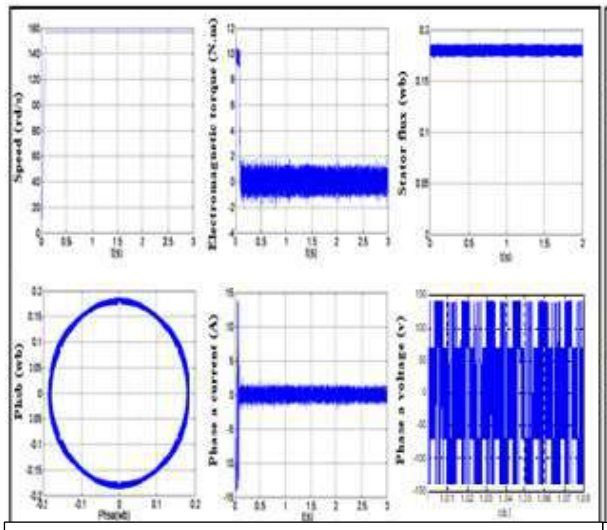


Figure (5.a): Dynamic behavior of the PMSM at no load

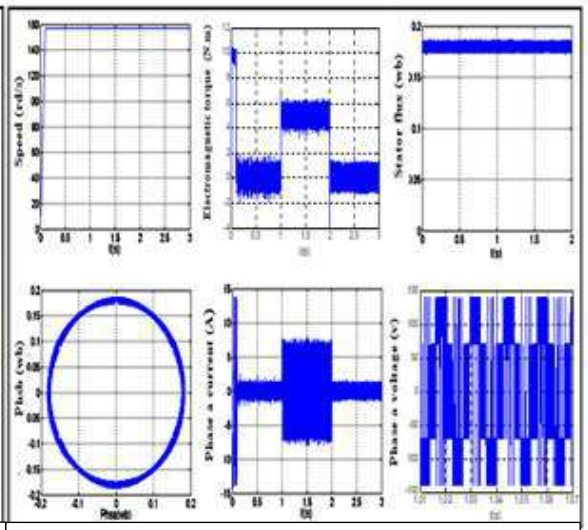


Figure (5.b): Dynamic behavior of the PMSM during startup under load variation

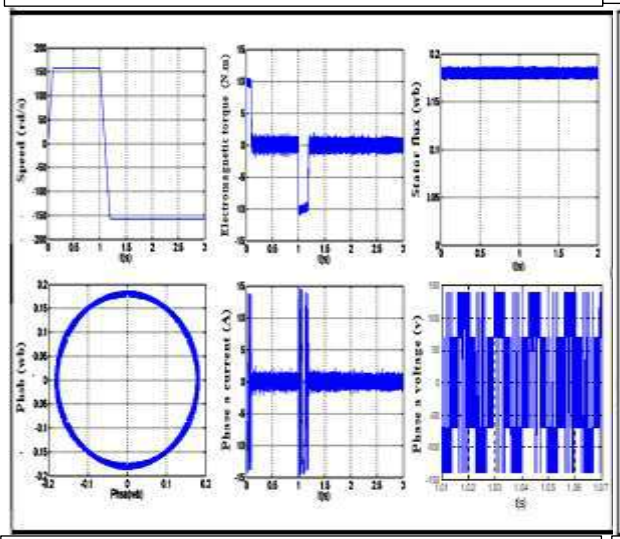


Figure (5.c): Dynamic behavior of the PMSM during a reversal of the direction of rotation.

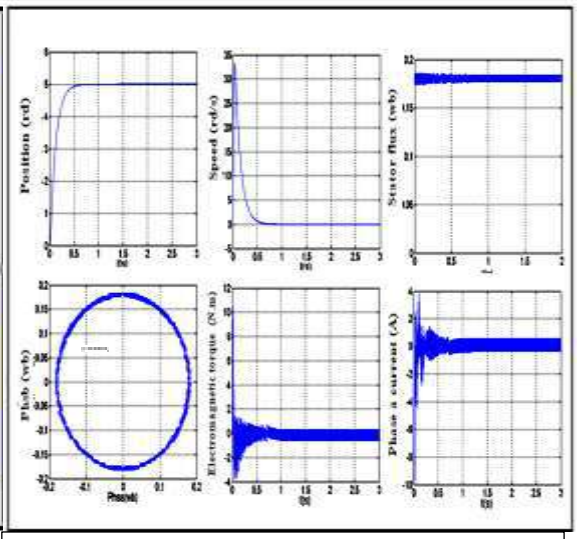


Figure (6.a): Dynamic behavior of the PMSM during no-load positioning

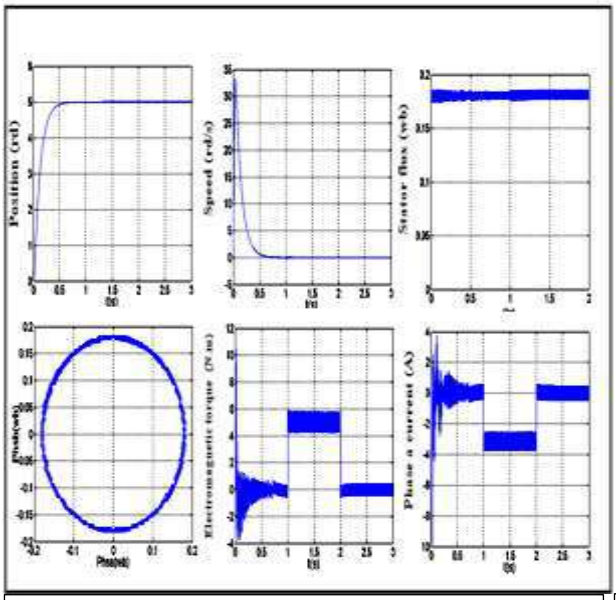


Figure (6.b): Dynamic behavior of the PMSM during positioning with load variation

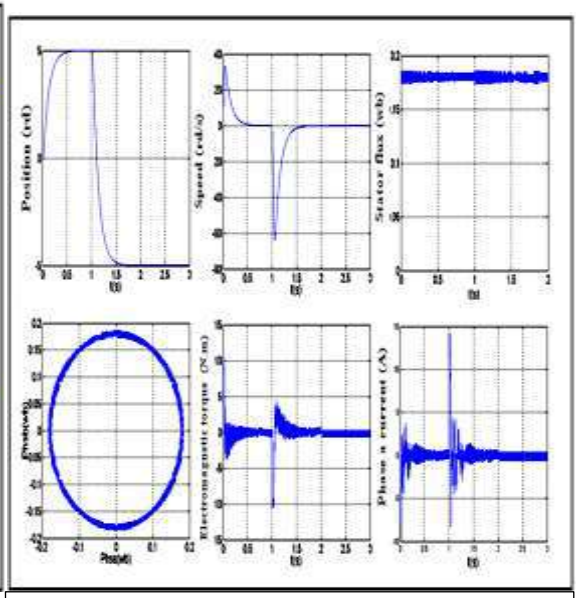


Figure (6.c): Dynamic behavior of the PMSM during position reversal

Source: Authors, (2026).

III.1.2 Indirect Identification-Based Fuzzy Adaptive Control

In this section, we propose a novel approach to indirect adaptive fuzzy control, combining fuzzy systems with the so-called computed torque method. Fuzzy systems are employed to provide a representative model of the direct dynamics of the system to be controlled. The control law is then derived from this model using the computed torque method. [13]

III.1.2.1 System Identification Using Fuzzy Systems

Consider again the system described by equation (3).

$$\begin{aligned} x_i^{(n)} &= F_i(X) + G_i(X)u_i \\ y_i &= x_i; \quad i = 1,\ldots,m \end{aligned} \tag{36}$$

The objective of the proposed control strategy is to generate a local control law based on (36) (a decentralized control structure), using estimates of the functions. F_i and $G_i$the control synthesis is performed in two steps:

- System identification: The model of the system to be controlled is identified using fuzzy systems.
- Control computation: The control law is derived via the computed torque method using the identified fuzzy model.

From Equation (40), we can write:

$$\hat{x}_i^{(n)} = f_i(X;\beta_{fi}) + g_i(X;\beta_{gi})u_i \tag{37}$$

The functions $f(.)_i \; and \; g_i(.)$ are first-order Sugeno fuzzy systems with linguistic descriptions of the following form: [24]

- For fuzzy system $f_i(\cdot)$:

$$\begin{aligned} &R_K^f : si\ x_l^{(n-1)}\ est\ F_{l(l,n-1)}^{\mathbf{f}}\ et..et\ x_m^{(n-1)}\ est\ F_{l(m,n-1)}^{f}\ et..et\ x_i\ est\ F_{l(1,0)}^{\mathbf{f}}\ et\ x_m\ est\ F_{l(m,0)}^{f} \\ &Alors\ \ s_f(k) = a_f(o,k) + a_f(1,n-1,k)x_1^{(n-1)} + \ldots + a_f(m,n-1,k)x^{n-1}{}_m + \ldots + a_f(1,0,k)x_1 \ldots \\ &\qquad + a_f(m,0,k)x_m \end{aligned} \tag{38}$$

- For fuzzy system : $g(\cdot)$:

$$\begin{aligned} &R_K^g : si\ x_l^{(n-1)}\ est\ F_{l(l,n-1)}^{\mathbf{g}}\ et..et\ x_m^{(n-1)}\ est\ F_{l(m,n-1)}^{g}\ et..et\ x_i\ est\ F_{l(1,0)}^{\mathbf{g}}\ et\ x_m\ est\ F_{l(m,0)}^{g} \\ &Alors\ \ s_g(k) = a_g(o,k) + a_g(1,n-1,k)x_1^{(n-1)} + \ldots + a_g(m,n-1,k)x^{n-1}{}_m + \ldots + a_g(1,0,k)x_1 \ldots \\ &\qquad + a_g(m,0,k)x_m \end{aligned} \tag{39}$$

The final outputs of the fuzzy systems $f_i(.) \; and \; g_i(.)$ are respectively given by:

$$f_i(X;\beta_{fi}) = \frac{\sum_{k=1}^{M_{fi}} \mu_{fk} s_f(K)}{\sum_{k=1}^{M_{fi}} \mu_{fk}} \tag{40}$$

$$g_i(X;\beta_{gi}) = \frac{\sum_{k=1}^{M_{gi}} \mu_{gk} s_g(k)}{\sum_{k=1}^{M_{gi}} \mu_{gk}} \tag{41}$$

Where:

$$\mu_{fK} = \mu_{l(1,n-1)}^{f} \times \mu_{Fl(2,n-1)}^{f} \times \ldots \times \mu_{Fl(m-1,0)}^{f} \times \mu_{Fl(m,0)}^{f} \tag{42}$$

$$\mu_{gK} = \mu_{l(1,n-1)}^{g} \times \mu_{Fl(2,n-1)}^{g} \times \ldots \times \mu_{Fl(m-1,0)}^{g} \times \mu_{Fl(m,0)}^{g} \tag{43}$$

The unknown parameters $\beta_i = \lfloor \beta_{fi}, \beta_{gi} \rfloor$ are estimated recursively using a modified gradient learning algorithm. In this algorithm, the parameters β_i are adjusted to minimize the instantaneous error between. $y_i^{(n)} et \; \hat{y}_i^{(n)}$ This error is given by: [35]

$$e_i(t) = y_i^{(n)} - \hat{y}_i^{(n)} \tag{44}$$

The proposed adaptation law takes the form [22]:

$$\hat{\beta}_i(t) = \hat{\beta}_i(t-1) + p(t)\psi(t)e_i(t)$$
$$\psi(t) = \frac{\partial y_i^{(n)}}{\partial \beta_i} = \left[\frac{\partial f_i(\cdot)}{\partial \beta_{fi}}; \frac{\partial g(\cdot)}{\partial \beta_{gi}}\right] \quad (45)$$

Where:

$$p(t) = \frac{\alpha_1 I}{\alpha_2 + \psi^T(t)\psi(t)}; \alpha_1 > 0, \alpha_2 > 0$$

III.1.2.2 Control Law Computation

Our objective is to design a control law such that each subsystem tracks its reference y_{id}. For this purpose, the control law, computed via the computed torque method, is given by:

$$u_i = \frac{v_i - f_i(X;\beta_{fi})}{g_i(X;\beta_{gi})} \quad (46)$$

Where:

$$v_i = y_{di}^{(n)} + k_n\left(y_{di}^{(n-1)} - y_i^{(n-1)}\right) + \ldots + k_1(y_{di} - y_i)$$

The coefficients k_i are selected such that the polynomial $s^n + k_n s^{n-1} \ldots + k_1$ is a Hurwitz polynomial (roots with negative real parts). This control law ensures critical damping for the error dynamics. Figure (7) schematically illustrates the principle of the proposed control strategy. [35].

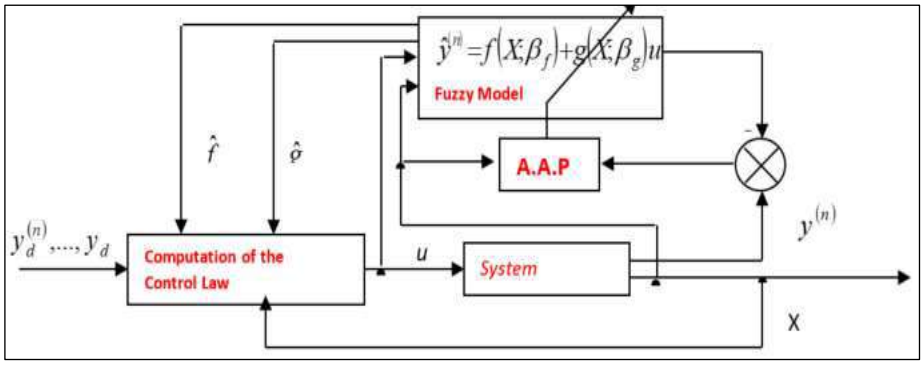


Figure 7: Principle of indirect adaptive fuzzy linearizing control.
Source: Authors, (2026).

✓ **Application to PMSM (Permanent Magnet Synchronous Motor)**

a) Speed Regulation

The structure of the indirect adaptive fuzzy linearizing control is illustrated in Figure (8). The adaptive fuzzy controller generates the current i_{qref}, which is multiplied by a coefficient $\left(\frac{3p\emptyset_f}{2}\right)$ to produce the electromagnetic torque T_{emref}. This torque, along with the reference flux Φ_{ref}, are then utilized in the Direct Torque Control (DTC) scheme to generate the inverter switching commands.

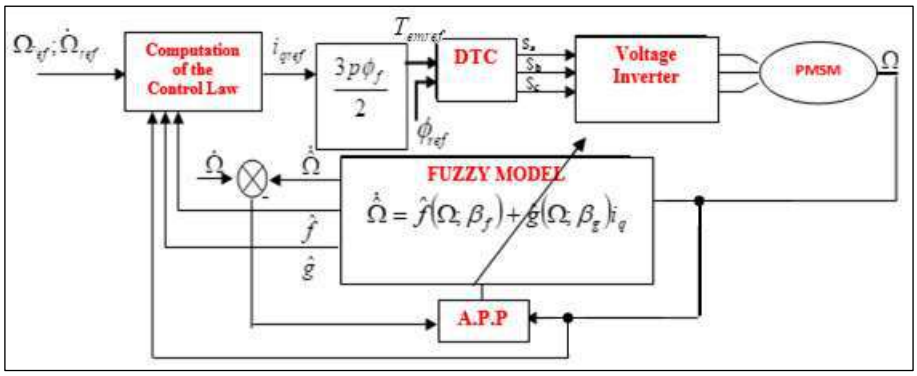


Figure 8: Speed control structure using the indirect adaptive fuzzy linearizing control method.
Source: Authors, (2026).

Considering equation (36), the selected model for identifying the speed derivative has the following form:

$$\dot{\Omega} = f(\Omega;\theta_f) + g(\Omega;\theta_g) i_q \tag{47}$$

With: $f(.) and\ g(.)$ are fuzzy systems with unknown parameters. These systems are characterized by a single input, with three membership functions assigned to this input. [35]. The final outputs of the fuzzy systems $f(.) et\ g(.)$ are respectively given by:

$$f(\Omega;\beta_f) = \frac{\sum_{k=1}^{3} \mu_{fk} s_f(k)}{\sum_{k=1}^{3} \mu_{fk}}$$
$$g(\Omega;\beta_g) = \frac{\sum_{k=1}^{3} \mu_{gk} s_g(k)}{\sum_{k=1}^{3} \mu_{gk}} \tag{48}$$

Where: $\mu_{fk} and\ \mu_{gk}$ represent the confidence degrees of rules $R_{fk} and\ R_{gk}$ for the fuzzy systems $f(.) and\ g(.)$. While $s_f(k) and\ s_g(k)$ are the consequences of these rules, defined as:

$$s_f(k) = a_f(0,k) + a_f(1,k)\Omega$$
$$s_g(k) = a_g(0,k) + a_g(1,k)\Omega \tag{49}$$

The output provided by the adaptive fuzzy controller is expressed as:

$$i_{qref} = \frac{v - f(\Omega;\beta_f)}{g(\Omega;\beta_g)} \tag{50}$$

With:

$$v = \dot{\Omega}_{ref} + k\Omega(\Omega_{ref} - \Omega)$$

From this, we obtain the reference electromagnetic torque through the equation:

$$T_{emref} = \frac{3p\phi_f}{2} \frac{v - f(\Omega;\beta_f)}{g(\Omega;\beta_g)} \tag{51}$$

b) Position Control

The control structure for position regulation is illustrated in Figure (9).

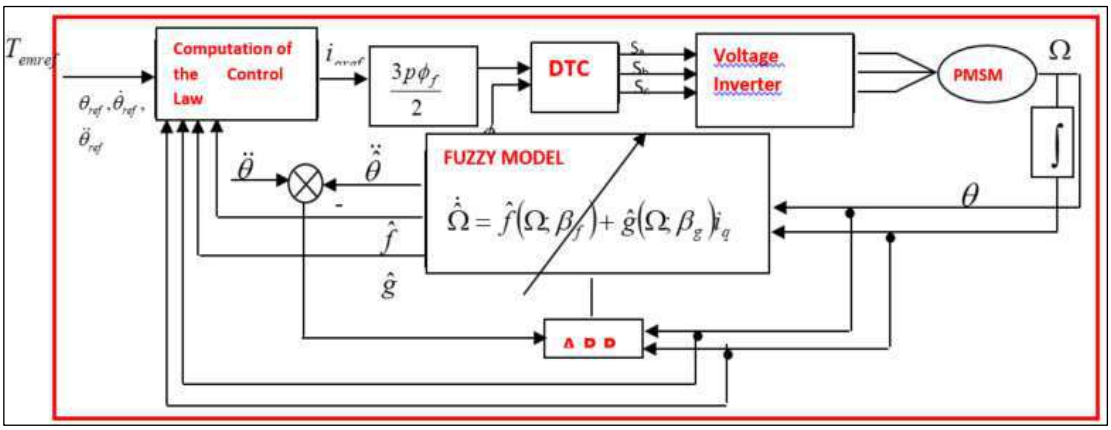


Figure 9: Position control architecture using indirect adaptive fuzzy linearization control methodology.
Source: Authors, (2026).

Based on equation (28), the selected identification model for $\ddot{\theta}$ is formulated as:

$$\ddot{\theta} = f(\ddot{\theta};\beta_f) + g(\ddot{\theta};\beta_g) i_q \tag{52}$$

Where $f(\ddot{\theta};\beta_f)\ and\ g(\ddot{\theta};\beta_g)$ represent fuzzy inference systems characterized by a single input variable. In the present study, we implement three Gaussian membership functions for this input. The aggregated outputs of these fuzzy systems are mathematically expressed as: [36]

$$f(\dot{\theta},\beta_f)=\frac{\sum_{k=1}^{3}\mu_{fk}s_f(k)}{\sum_{k=1}^{3}\mu_{fk}}$$
$$g(\dot{\theta},\beta_g)=\frac{\sum_{k=1}^{3}\mu_{gk}s_g(k)}{\sum_{k=1}^{3}\mu_{gk}} \quad (53)$$

The control signal generated by the adaptive fuzzy controller is given by:

$$i_{qref}=\frac{v-f(\dot{\theta},\beta_f)}{g(\dot{\theta},\beta_g)} \quad (54)$$

With:

$$v=\ddot{\theta}_{ref}+k_{1\theta}(\dot{\theta}_{ref}-\dot{\theta})+k_{2\theta}(\theta_{ref}-\theta)$$

The reference electromagnetic torque is consequently derived from:

$$T_{emref}=\frac{3p\phi_f}{2}\frac{v-f(\dot{\theta},\beta_f)}{g(\dot{\theta},\beta_g)} \quad (55)$$

Remark 1: The adaptation algorithm necessitates real-time measurement of the angular velocity derivative. While numerically estimated in this implementation, we note that such approximation may introduce significant errors under noisy measurement conditions, potentially c]ompromising control performance.

- Numerical Simulation and Performance Analysis

The closed-loop system dynamics were rigorously evaluated through numerical simulations investigating both speed and position regulation using the proposed indirect adaptive fuzzy linearization approach. The critical control parameters, determining the PMSM's dynamic response characteristics, were optimally tuned to 620.54, 61.5704, and 473.86 respectively. The transient and steady state performance metrics, illustrated in Figures (10, a-c) and (11, a-c) demonstrate:

- Exceptional disturbance rejection capability:
- Maximum speed deviation: 0.55% of nominal value
- Maximum position error: 0.46% of reference
- Perturbation recovery time constants: 5ms (speed) and 290ms (position)
- Robust flux-torque decoupling
- Complete independence maintained even under severe operational regimes
- No observable coupling effects during rapid torque transients

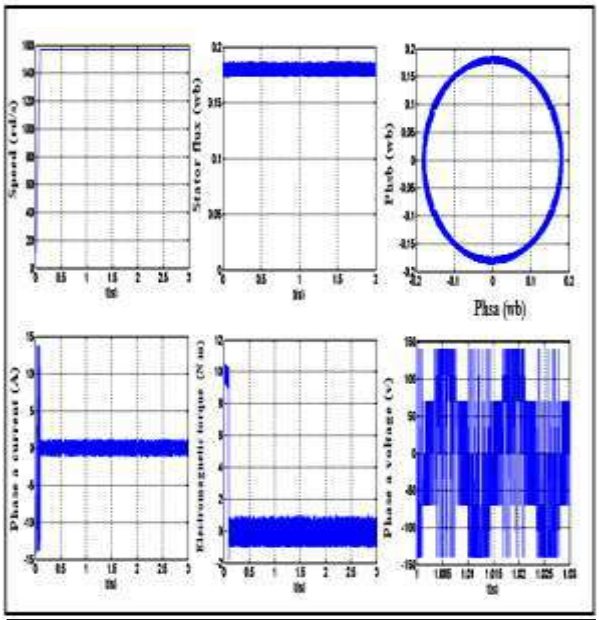


Figure (10.a): Dynamic behavior of the PMSM under no-load conditions

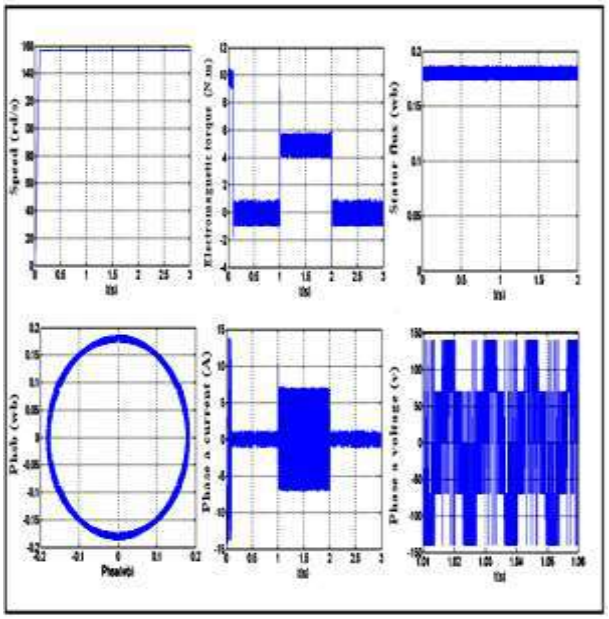


Figure (10.b): Dynamic response of the PMSM during startup under variable load conditions

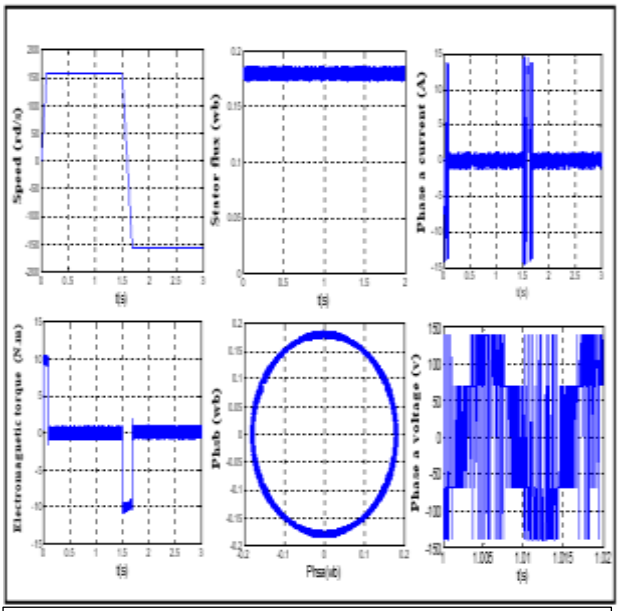


Figure (10.c): Dynamic behavior of the PMSM during rotational direction reversal

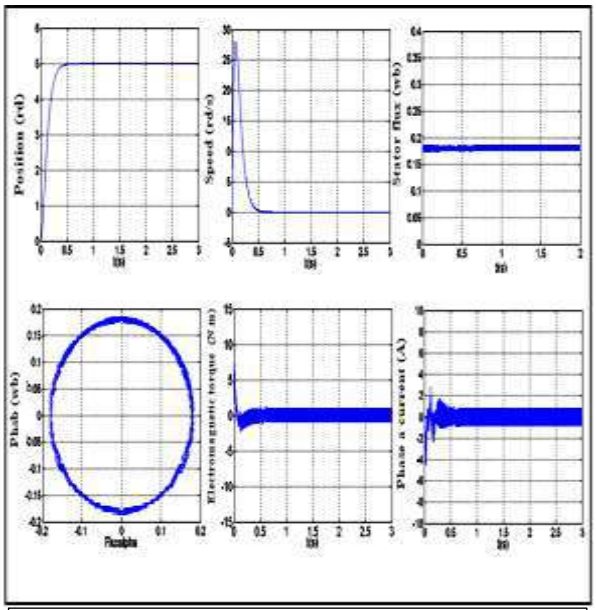


Figure (11.a): Dynamic positioning performance of the PMSM under no-load conditions

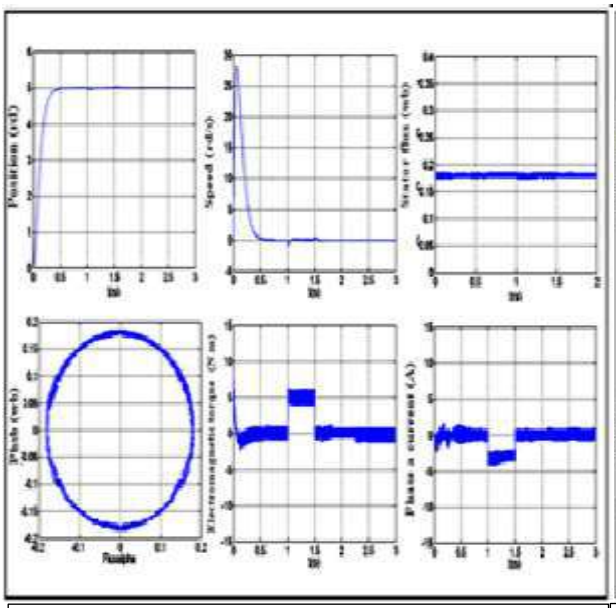


Figure (11.b): Dynamic positioning performance of the PMSM under variable load conditions

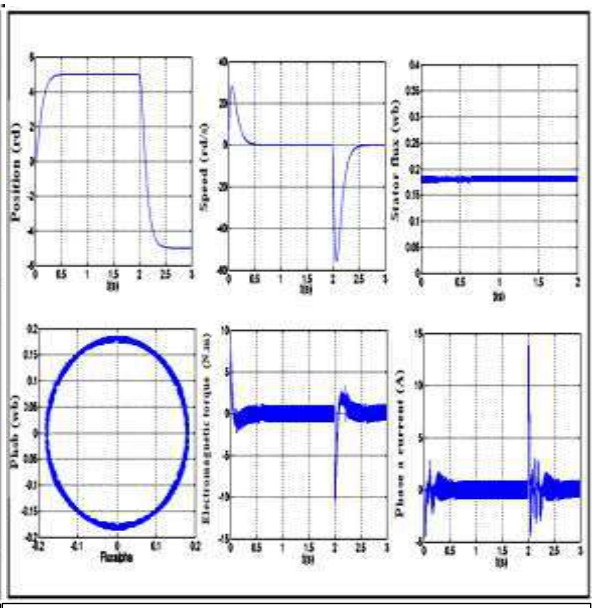


Figure (11.c): Dynamic positioning behavior of the PMSM during position reversal

Source: Authors, (2026).

IV. CONCLUSIONS

This research has systematically investigated two adaptive fuzzy control approaches for permanent magnet synchronous motors (PMSMs): direct and indirect adaptive fuzzy control methodologies. The study demonstrates how fuzzy systems can effectively approximate unknown continuous nonlinear functions in both control paradigms, while maintaining complete independence from prior knowledge of the PMSM's dynamic model structure or parameters. Our contributions include:

Direct Adaptive Fuzzy Control: A novel control architecture where fuzzy logic systems directly generate control signals through online adaptation.
Indirect Adaptive Fuzzy Linearization Control: An identification-based approach employing fuzzy systems for model approximation and linearization
Both decentralized control strategies exhibit:
Robust position and speed regulation capabilities
Model-free operation without requiring parametric knowledge
Fuzzy rule bases derived directly from the corresponding fuzzy models
Consistent operation across various dynamic regimes

However, these techniques require the measurement of the machine's speed derivative, which is calculated numerically, which can be detrimental in the case of noisy measurements. The study establishes a theoretical framework for advanced fuzzy control applications in PMSM drives, particularly suitable for industrial applications requiring both precision and robustness under uncertain operating conditions. Future work should focus on experimental validation and computational optimization for real-time implementation.

V.AUTHOR'S CONTRIBUTION

Conceptualization: Bentchikou Ibrahim, Tlemçani Abdelhalim, Hassan Nouri, Boudjema Fares, Boukhetala Djamel,Ould cherchali Noureddine ,Fekir Mohamed, Mahdab Salim.
Methodology: Bentchikou Ibrahim, Tlemçani Abdelhalim, Hassan Nouri, Boudjema Fares, Boukhetala Djamel,Ould cherchali Noureddine ,Fekir Mohamed, Mahdab Salim .
Investigation: Bentchikou Ibrahim, Tlemçani Abdelhalim, Hassan Nouri, Boudjema Fares, Boukhetala Djamel,Ould cherchali Noureddine ,Fekir Mohamed, Mahdab Salim.
Discussion of results : Bentchikou Ibrahim, Tlemçani Abdelhalim, Hassan Nouri, Boudjema Fares, Boukhetala Djamel,Ould cherchali Noureddine ,Fekir Mohamed, Mahdab Salim.
Writing–Original Draft: Bentchikou Ibrahim, Tlemçani Abdelhalim, Hassan Nouri, Boudjema Fares, Boukhetala Djamel,Ould cherchali Noureddin ,Fekir Mohamed,Mahdab Salim.
Writing –Review and Editing: Bentchikou Ibrahim, Tlemçani Abdelhali, Hassan Nouri, Boudjema Fares, Boukhetala Djamel,Ould cherchali Noure ,Fekir Mohamed, Mahdab Salim.
Resources: Bentchikou Ibrahim, Tlemçani Abdelhalim, Hassan Nouri, Boudjema Fares, Boukhetala Djamel,Ould cherchali Noureddine ,Fekir Mohamed, Mahdab Salim .
Supervision: Bentchikou Ibrahim, Tlemçani Abdelhalim, Hassan Nouri, Boudjema Fares, Boukhetala Djamel,Ould cherchali Noureddine ,Fekir Mohamed, Mahdab Salim .
Approval of the final text: Bentchikou Ibrahim, Tlemçani Abdelhalim, Hassan Nouri, Boudjema Fares, Boukhetala Djamel,Ould cherchali Noureddi ,Fekir Mohamed, Mahdab Salim.

VI. REFERENCES

[1] I,.Bentchikou; K, Halbaoui.; F, Boudjema.; D,Boukhetala and A, Tlemçani.“ Alternative hybrid control of switched systems. an application to machine dc fed by multicellular converter”, Revue Roumaine des Sciences Techniques, Série Électrotechnique et Énergétique, Vol. 67, 3, pp. 247–252, Bucarest, 2022, ISSN 0035-4066.

[2] V. Sri Charan, D. Kiran Kumar, “speed control of pmsm drive using model predictive control based field oriented control:”, JETIA, v.11 n.52, p. 1 -8.March/April, 2025

[3] A. N. Halisyah, D. Adiputra, and F. A. Al. "Field oriented control driver development based on BTS7960 for physiotherapy robot implementation". International Journal of Electrical and Computer Engineering (IJECE), vol. 14, no. 2, pp. 1486–1495, 2024. DOI: 10.11591/ijece.v14i2.pp1486-1495.

[4] E. Benmalek, M. Rayyam, A. Gege, O. Ennasiri, and A. Ezzaidi. "AI-BASED FIELD-ORIENTED CONTROL FOR INDUCTION MOTORS". Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska, vol. 14, no. 4, pp. 75–81, 2024. DOI: 10.35784/iapgos.6253.

[5] T. Le. "A Robust Control Strategy for Effective Field-Oriented Control of PMSMs". Engineering, Technology & Applied Science Research, vol. 14, no. 6, pp. 18469–18475, 2024. DOI: 10.48084/etasr.8893.

[6] M. P. Kazmierkowski; L. Malesani; L. Zarri et A. Tani. "Direct Torque Control of PWM Inverter-Fed AC Motors - Survey and Recent Advances", IEEE Transactions on Power Electronics, vol. 38, no. 1, pp. 1-20, 2024. DOI:10.1109/TPEL.2023.3304561.

[7] S. Vazquez; J. I. Leon; L. G. Franquelo et M. Reyes. "Predictive Torque Control for High-Performance Drives: A Comparative Study with DTC and FOC", IEEE Transactions on Industrial Electronics, vol. 71, no. 3, pp. 2105-2118, 2025. DOI:10.1109/TIE.2024.3356789.

[8] M. Ahmed et al. "DTC flou adaptatif pour PMSM", IEEE Trans. on Transportation Electrification, vol. 10, no. 1, 2024. DOI: 10.1109/TTE.2023.3304561.

[9] M. Ogharandukun, F. I. Onah, N. B. Ngang et E. Chijioke. "Adaptive Control-Based Technique for Enhancing Energy Efficiency of a Wind Energy Generator". European Journal of Engineering and Environmental Sciences, vol. 8, no. 1, pp. 21–29, 2024. DOI: 10.5281/zenodo.12787161.

[10] V. Stepanyan et K. Krishnakumar. "Adaptive Control with Reference Model Modification". Journal of Guidance, Control, and Dynamics, vol. 35, no. 4, pp. 1370–1374, 2012. DOI: 10.2514/1.55756.

[11] T. Yucelen et A. J. Calise. "Derivative-Free Model Reference Adaptive Control". Journal of Guidance, Control, and Dynamics, vol. 34, no. 4, pp. 933–950, 2011. DOI: 10.2514/1.53234.

[12] K. J. Åström and T. Hägglund. "The future of PID control". Control Engineering Practice, Vol. 9, No. 11, pp. 1163–1175, 2001. DOI: 10.1016/S0967-0661(01)00062

[13] T. Jin, X. Yang, C. Qi, Y. Wu, Y. Zhang and W. Xu. "Model reference adaptive PID control for enhanced stability and posture control of unmanned underwater vehicles". Ocean Engineering, Vol. 330, 121218, 2025. DOI: 10.1016/j.oceaneng.2025.121218.

[14] N. Essounbouli. "Contrôleur flou adaptatif pour robots mobiles", JofeSys vol. 21, no. 1, 2025.

[15] N. Aliman, R. Ramli, S.M. Haris, M.S. Amiri, M. Van, A robust adaptive-fuzzyproportional-derivative controller for a rehabilitation lower limb exoskeleton, EngSci. Technol. Int. J. 35 (2022) 101097.

[16] W. Xiao, K. Chen, J. Fan, Y. Hou, W. Kong, G. Dan, AI-driven rehabilitation and assistive robotic system with intelligent PID controller based on RBF neural networks, Neural Comput. Appl. 35 (22) (2023) 16021–16035.

[17] M. Gheisarnejad, M.H. Khooban, An intelligent non-integer PID controller-based deep reinforcement learning: implementation and experimental results, IEEE TransInd. Electron. 68 (4) (2020) 3609–3618.

[18] J. Hu, H. Zhang, H. Liu, X. Yu, A survey on sliding mode control for networked control systems, Int. J. Syst. Sci. 52 (6) (2021) 1129–1147.

[19] A.G. Loukianov, J.R. Dominguez, B. Castillo-Toledo, Robust sliding mode regulation of nonlinear systems, Automatica 89 (2018) 241–246

[20] Y. Hu, J. Meng, G. Li, D. Zhao, G. Feng, G. Zuo, et al., Fuzzy adaptive passive control strategy design for upper-limb end-effector rehabilitation robot, Sensors 23 (8) (2023) 4042.

[21] H. Yang, Y. Jiang, S. Yin, Adaptive fuzzy fault-tolerant control for Markov jump systems with additive and multiplicative actuator faults, IEEE Trans. Fuzzy Syst. 29 (4) (2020) 772–785.

[22] M. Munadi, B. Radityo, M. Ariyanto, Y. Taniai, Automated guided vehicle (AGV) lane-keeping assist based on computer vision, and fuzzy logic control under varying light intensity, Results Eng. 21 (2024) 101678. Results in Engineering 24 (2024) 102916 13

[23] V.T. Nguyen, H.B. Giap, N.T. Tran, N.H. Manh, V.A. Nguyen, A study on fuzzy nonsingular fast terminal sliding mode control for pendubot with uncertainties, in: Springer ICISN, 2023, pp. 12–18.

[24] F. Wang, B. Chen, Y. Sun, Y. Gao, C. Lin, Finite-time fuzzy control of stochastic nonlinear systems, IEEE Trans. Cybern. 50 (6) (2019) 2617–2626.

[25] Y. Pan, P. Du, H. Xue, H.K. Lam, Singularity-free fixed-time fuzzy control for roboticsystems with user-defined performance, IEEE Trans. Fuzzy Syst. 29 (8) (2020) 2388–2398.

[26] X. Su, F. Xia, J. Liu, L. Wu, Event-triggered fuzzy control of nonlinear systems with its application to inverted pendulum systems, Automatica 94 (2018) 236–248.

[27] H. Zhang, Y. Liu, Y. Wang, Observer-based finite-time adaptive fuzzy control for nontriangular nonlinear systems with full-state constraints, IEEE Trans. Cybern. 51 (3) (2020) 1110–1120.

[28] M. Van, Y. Sun, S. Mcllvanna, M.N. Nguyen, M.O. Khyam, D. Ceglarek, Adaptive fuzzy fault tolerant control for robot manipulators with fixed-time convergence, IEEE Trans. Fuzzy Syst. 31 (9) (2023) 3210–3219.

[29] V.T. Nguyen, Q.C. Nguyen, M.Van, V.T. Nguyen, D.H.Pham, & D.C Nguyen, ‘’Novel adaptive fuzzy control for pendubot with actuator faults and uncertainties: design and experiments’’ Results in Engineering, 24, Article 102916 ://doi.org/10.1016/j.rineng.2024.102916

[30] L.Daniel , S.Igor , G.Fernando ‘’From Model-Based and Adaptive Control to Evolving Fuzzy Control’’2025 IEEE International Conference on Fuzzy Systems

[31] I.Bentchikou, F. Boudjema, D. Boukhetala ,A.Tlemçani, N. Ould Cherchali “ Investigation in the Technique of Adaptive Predictive Control Fed by a Hybrid Inverter Applied to a Permanent Magnetic Synchronous Machine”, Nonlinear Dynamics and Systems Theory. An International Journal of Research and Surveys . Volume 16, Number 1, 2016. ISSN 1562-8353.

[32] A. Tlemçani; F. Boudjema; D. Boukhetala. "Nonlinear Backstepping Control Strategy for PMSM with Uncertain Load Torque", International Journal of Electrical Power & Energy Systems, vol. 155, 109567, 2024.

[33] P. Du, K. Sun, S. Zhao, and H. Liang, ‘‘Observer-based adaptive fuzzy control for time-varying state constrained strict-feedback nonlinear systems with dead-zone,’’ Int. J. Fuzzy Syst., vol. 21, no. 3, pp. 733–744, Apr. 2019.

[34] S. Ha, L. Chen, H. Liu, and S. Zhang, ‘‘Command filtered adaptive fuzzy control of fractional-order nonlinear systems,’’ Eur. J. Control, vol. 63, pp. 48–60, Jan. 2022.

[35] X. You, S. Dian, K. Liu, B. Guo, G. Xiang, and Y. Zhu, ‘‘Command filter-based adaptive fuzzy finite-time tracking control for uncertain fractional-order nonlinear systems,’’ IEEE Trans. Fuzzy Syst., vol. 31, no. 1, pp. 226–240, Jan. 2023.

[36] A. Bounemeur and M. Chemachema, ‘‘Finite-time output-feedback fault tolerant adaptive fuzzy control framework for a class of MIMO saturated nonlinear systems,’’ Int. J. Syst. Sci., vol. 56, no. 4, pp. 733–752, Mar. 2025.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1197-1206. January/February, 2026.

DOI: https://doi.org/10.5935/jetia. v12i57.2741

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

PARTICLE SWARM OPTIMIZATION AND ENSEMBLE METHODS FOR ABSENTEEISM PREDICTION

Baydaa Hadi Saoudi[1], Tabarek Alwan Tuib[*2] and Yaqdhan Mahmood Hussein[3]

[1]Technical Institute-Samawa, Information Technology and Communication (ICT), Al-Furat Al-Awast Technical University,Samawa, Iraq.
[2, 3]Collage of Engineering, Electronics and Communications Engineering, Al Muthanna University, Muthanna, Iraq
[3]Al Ayen University, Nasiriyah 64001, Iraq

[1]https://orcid.org/0000-0002-8198-324, [2]https://orcid.org/0000-0002-1646-0015, [3]https://orcid.org/0000-0001-7153-4120

Email: ins.byd44@atu.edu.iq, *tabarcka@gmail.com, Yaqthanm79@gmail.com.

ARTICLE INFO

Article History
Received: September 25, 2025
Revised: October 20, 2025
Accepted: December 15, 2026
Published: February 28, 2026

Keywords:
Absenteeism,
Particle Swarm Optimization,
Ensemble classification.

ABSTRACT

As it might lead to the loss of productivity and an economy, high absenteeism of employees badly affects organizations. With regard to courier companies, the methods of machine learning (ML) are utilized for the absenteeism of employees. The research uses inductive learning method, using conventional methods of data mining (DM), like data preprocessing, feature selection (FS) by Particle Swarm Optimization (PSO), and data classification by using such ensemble techniques, like Boosting, Voting, Bagging, and Stacking. The dataset with 21 attributes were 740 instances and this was processed to eliminate irrelevant features and transformed for effective analysis. The proposed model was very accurate with Boosting being the most successful ranking at 99.8 accuracy. The research indicates that FS is significant in enhancing accuracy of predictions and it is a useful instrument that managers can use to reduce the effects of absenteeism in organizational performance. In future, it will be possible to incorporate additional datasets, investigate more advanced methods of FS and to incorporate external variables to improve prediction.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

ML is regarded as a pivotal domain in research, aimed in uncovering or forecasting subjects [1]. Artificial intelligence enables machines to exhibit cognitive functions. Furthermore, artificial intelligence enhances the cognitive capabilities of machines. ML constitutes a domain within the field of artificial intelligence study.[2] Certain researchers assert that intelligence cannot be cultivated without education. [3] All of these methods have been used for categorizing the subjects of dataset. In supervised learning, a training set of examples with corresponding targets is supplied, and algorithms responding properly to all the potential inputs depending on this training set. Supervised learning encompasses classification and regression. [4] The semi-supervised learning methodology can be defined as a variant of supervised learning. In addition, unlabeled data was used for the training task in this approach. [5] Furthermore, DL is a part of ML built on a set of algorithms. [6] There is an urgent need for ML approaches in today's environment in order to reduce human effort while reducing errors and achieving the greatest degree of automation.[7] Absenteeism is widely seen as a significant signal of inadequate performance inside a business.[8] Absenteeism from work has negative impact on the country's gross domestic product.

Furthermore, the annual expense of absenteeism in nations is approximated to be between 1.2 and 2% of GDP.[9] Absenteeism among blue-collar workers may stem from several factors, including religious and social commitments, inadequate housing, also subpar working circumstances. Organizations employ several strategies to mitigate absenteeism, such as adopting work rotation and enrichment, providing financial incentives and bonuses, and enforcing punctuality regulations. [10] Absenteeism is a pertinent issue in human resource management across many industries and organizations [11], [12] in the context of international human resource management (IHRM). The Micro-HR management can be defined as the sub-functions regarding HR policy and it has two main categories: the first one is related to the management of people and small groups, which involve the selection and recruitment, orientation and training, performance management, in addition to rewards. The other is work organization management and employee voice systems [13], [14]. The issue of absenteeism is a major obstacle which must be tackled by any company and employee.

The employees should have a sense of fulfillment in their employment and the employers should help them have this good experience in every possible way. Absenteeism is a common pattern of non-attendance of some responsibility or duty. The issue of absenteeism in employees is not purely a financial issue; it has a great effect on employee morale as well. Absenteeism in the workplace can be a pointer to low morale among the employees and management [15]. Therefore, one of the important issues to consider when making predictions of human behavior, especially absenteeism, is the means in which large datasets can be shifted or refined. The FS is an initial step in the process of DM. It aims at removing the extraneous and redundant features of the original data [16], [17]. In most cases, companies strive to increase their profits.

On-demand, casual, and temporary employment (which is common in the gig economy) can be used in organizations where employees are involved in simple tasks because of a lesser number of contractual obligations. The importance of employee specialization and job permanence are increased in companies where employees are doing specialized tasks. The importance of skills and knowledge has been defined and the ability to engage in learning continuously has been an urgent need to firms. The paper will be structured as mentioned: the second part clarifies a number of approaches and evaluates the relevant literature, whereas the third part outlines the data analysis, feature extraction, and CV classification techniques, and describes the proposed ML system. Part IV outlines the experimental results as well as the analysis of results. The fifth part compares the proposed model with the existing studies. The conclusion is marked in the end by Section VI.

II. THEORETICAL REFERENCE

Prior study endeavored to employ various data mining techniques or integrate distinct models to address the issue of absence prediction. Martiniano and colleagues. A neuro-fuzzy network was created using a multilayer perceptron design in an artificial neural network, implementing the error backpropagation method. This report offers initial findings from the utilization of the neuro-fuzzy network in forecasting workplace absence [18], [19]. A fuzzy neural network was constructed using error backpropagation algorithm with a multilayer perceptron to forecast absenteeism. Nonong and colleagues A study was undertaken to develop decision tree designed for detecting prevalent traits regarding employees who often exhibit absenteeism from work [20]. In a recent research, Gayathri employed naive Bayesian classifiers, multiclass perceptron, also J48 classifiers for developing classification model for predicting the employee absenteeism over extended or short durations [21].By [22] They employed artificial neural networks in their study.

Furthermore, in a pertinent study [23], researchers advocated for using neural networks (NNs) as well as DL algorithms for predicting employee commitment behaviors within the workplace. Conducting comprehensive research on large amount of data that is collected via businesses is a laborious and resource intensive task. Too much irrelevant information makes it hard to understand and comprehend [24]. As a result, effective filtering and narrowing down of large volumes of data is one of the focal points of human behavior prediction, especially when it comes to absenteeism. One of the most important preprocessing stages in DM is FS, which tries to remove irrelevant and redundant features of the original dataset [25], [26]. Addition of extraneous and redundant data into the model can lead to high expenditures, long time wastage, and reduction in the accuracy [27], [28]. The existing literature focuses on the lack of studies that use DM-based research to predict absenteeism and the lack of emphasis on the crucial aspect of FS.

By [29] noted that no studies aimed at the development of appropriate methods to compare and evaluate the effectiveness of different DM classification methods and the current literature also content highlight the lack of research using DM methods to predict absenteeism, which has led to the description of the absence of emphasis on the critical aspect of FS. Noted that there is no research aimed at developing adequate approaches to compare and evaluate the effectiveness regarding various DM classification methods and specific sets of features [29]. An alternative ML algorithm is applied in [30]. The study uses the linear regression and support vector regression (SVR) for the purpose of coming up with prediction model of absenteeism. Linear regression is performed with the use of all characteristics, and SVR is performed with consideration of two variables, namely, age and season. The results of the two algorithms are used to show that they are effective in predicting occupational absenteeism with high accuracy [31].

The proposed research will search and evaluate suitable ML tools to predict and evaluate workplace absenteeism. Parameters in the information being reviewed are age, education, employment category, day, month and term of service consisting of 125,000 records in total. The Weka tool is used to compare and contrast different algorithms in terms of accuracy, precision, and sensitivity [32]. This paper is an analysis of the absenteeism within courier company located in Brazil. Methods of ML were applied for understanding and prediction. It has already passed through preprocessing and different ML classification methods have been applied and these include ZeroR, J48 tree-based, k-NN, and NB classification methods. The study introduces the algorithms which could predict the absenteeism to accuracy of 92%. In addition, disciplinary failure has come out as one of the most predictive of absenteeism among the first 20 qualities [33].

The purpose of the presented work was discovering significant aspects as well as the best classification algorithms enhancing predicting absenteeism accuracy. We utilized FS methods for improving the efficiency regarding absenteeism prediction. The results of the experiment proved the importance of the selected features. Moreover, bagging has the best performance with accuracy of 92% in the seven prominent methods that apply CFS in FS [34]. Absenteeism is a major threat to businesses in the world. Our experimental group used logistic regression (LR), random forest (RF), XGBoost, and SVM for analyzing the data of 32 blue-collar employees within fast-moving consumer goods organization between 2017 and 2019. XGBoost and RF possessed F1 of 0.6399 and 0.6461, respectively, indicating fair predictive accuracy of absence. However, there are major growth prospects [35].

III. ADOPTED METHODOLOGY

III.1 PROPOSED METHOD

The research will attempt to formulate a predictive model of employee absenteeism in a courier company through inductive learning model. The model employs the conventional DM, which involves the application of Scikit-learn in Python. The model follows 3 stages including data preparation, FS with the help of PSO, and classification using ensemble methods.

It is aimed at precisely predicting absenteeism to help managers make decisions. This model applies preprocessing, PSO, bagging, boosting, voting, and stacking techniques in order to boost the predictivity. The performance related to integrated classification model is evaluated in terms of statistical metrics.

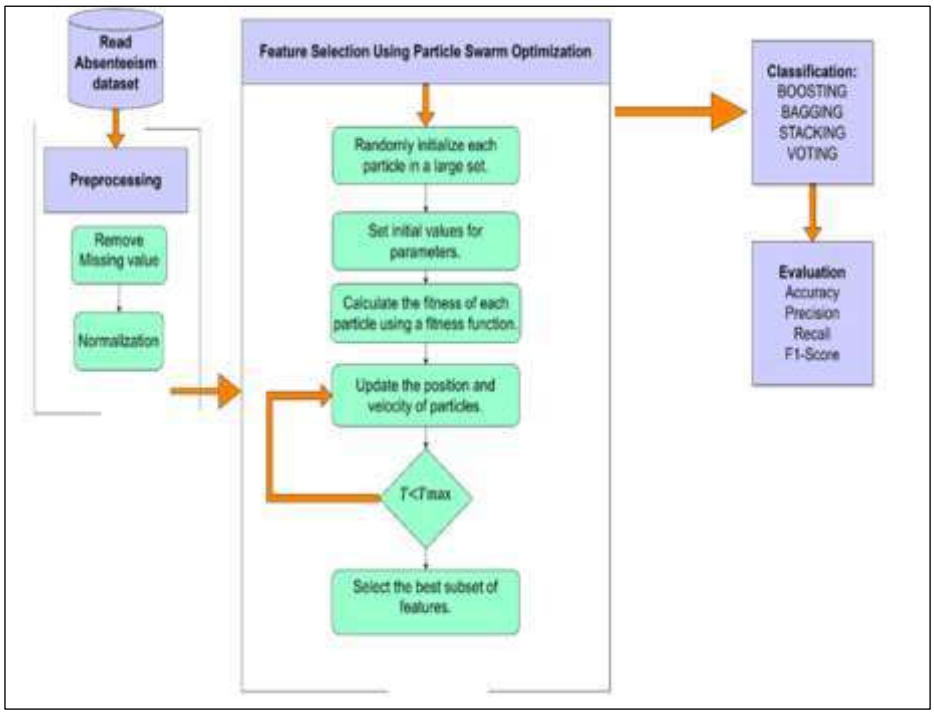


Figure 1: The proposed methodology for resume classification.
Source: Authors, (2026).

III.2 DATA SET

In this paper, we will examine absenteeism in courier company in Brazil using a dataset. The dataset is on Kaggle and records employees' attendance and cause of absenteeism among the employees, and other factors. It assists in determining trends and offers information on factors that have an impact on absenteeism. There are also work-related parameters such as work load and achievement targets. This dataset consists of 740 incidences of more than 21 attributes in three years. The paper proposes ML solutions to successful human resource strategies and employee interaction. The DL techniques, however, are not used because of its small size. Data is broken into training and test data.

III.3 FEATURE SELECTION USING PARTICLE SWARM OPTIMIZATION

III.3.1 Particle Swarm Optimization

In the PSO algorithm, each one of the particles is represented as a vector containing continuous values between 0 and 1. Each element in this vector represents a particular feature in the dataset. These continuous values are converted to binary values (0 or 1) using a binary conversion function. In the case when the value is greater than a specified threshold (0.5), it is assigned to 1 (feature is selected), and if it is less than the threshold, it is assigned to 0 (the feature is not selected).

III.3.2 Fitness Function

The objective function is the criterion by which the quality of each particle (solution) is evaluated. In this code, objective function is designed to evaluate the performance regarding a selected set of features based on two main criteria:

III.3.2.1 Error Rate

A classification model (such as Bagging, Boosting, Stacking and voting) is trained using the selected features from the particle. The error rate is then calculated depending on the model's performance on validation data. The goal is to minimize the error rate as much as possible.

III.3.2.2 Feature Size

The number of selected features is minimized to avoid overfitting and improve model efficiency. The goal is to select the fewest possible features while maintaining classification accuracy

$$\text{fitness} = \text{ErrorRate} * \alpha + \beta * \frac{\text{Number of Selected Features}}{\text{Total Features}} \quad (1)$$

The importance of each criterion is controlled using the parameters α and β, where α represents the weight of the error rate (0.99), prioritizing classification accuracy, while β = 1 - α represents the weight of the number of features (0.01), reducing the impact of the number of features compared to the error rate. The error rate is calculated based on the performance of the classifier, and the number of selected features is the number of features selected from the total number of features

III.4 ENSEMBLE CLASSIFICATION

III.4.1 Bagging

Bagging is implemented with the use of RF algorithm, where random samples are taken with the return from the training data. Multiple independent decision trees are trained on these samples, and their results are then combined utilizing majority voting in classification. The figure 2 illustrates how Bagging works, where random samples of data are taken, sub-models (RF) are trained in parallel on these samples, and then their results are combined to obtain the final prediction

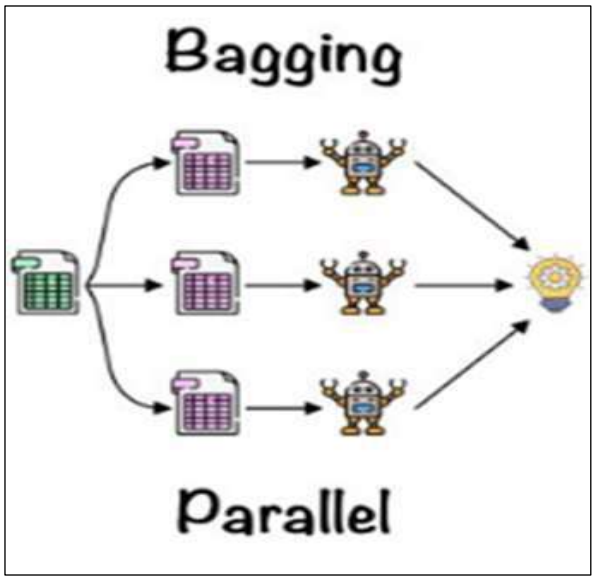


Figure 2: Bagging process.
Source: Authors, (2026).

III.4.2 Boosting

Boosting is based on the RF, and weak models are trained in sequence. The weight assigned to the samples that were misclassified at every step to enhance their performance in the successive models is higher. The outcome of the models is progressively increased in a way that enhances a stronger and more accurate classifier. The figure 3 depicts the Boosting process in the project whereby the models (RF) are trained in series, and all the models enhance the performance regarding other models upon the errors of preceding model and finally, the results are combined to acquire a strong classifier.

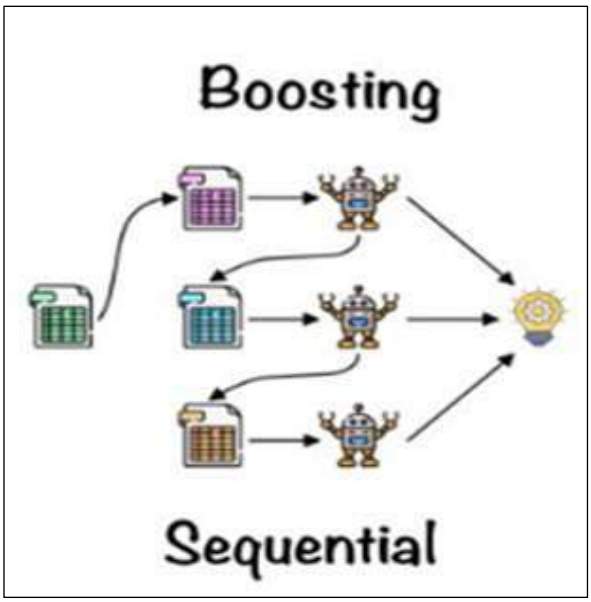


Figure 3: Boosting process.
Source: Authors, (2026).

III.4.3 Voting

The results of three basic models (LR, NB and SVM) were combined by voting in order to enhance the classification accuracy. This is the method by which hard voting is applied to the final decision which is made based on the principle of voting. Both models are applied to the same dataset and the final choice in the models is made based on the majority of the predictions of the models. The figure 4 presents the performance of three simple classifiers (LR, NB, and SVM) on the same dataset. The voting component is used to combine the results of these classifiers and make the final decision based on the majority.

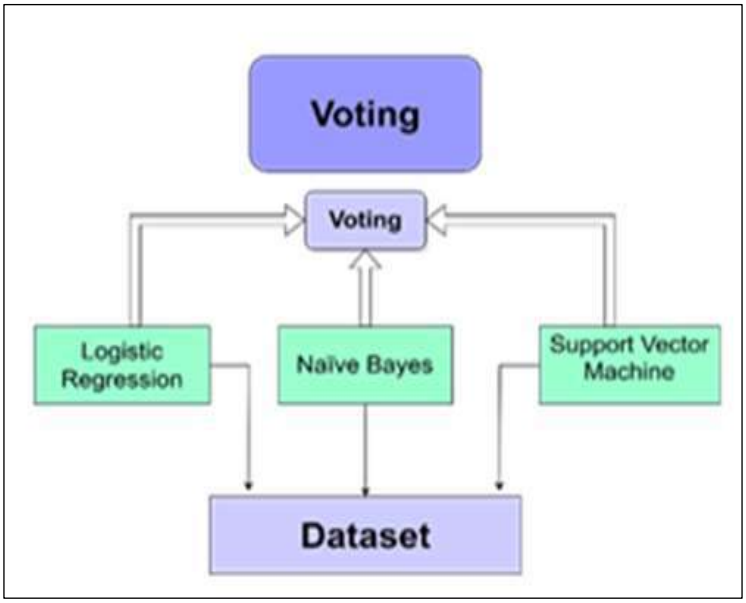


Figure 4: Voting process.
Source: Authors, (2026).

III.4.4 Stacking

The method is dependent on the training of a number of weak base models, including LR, k-NN, and SVM, on training data. The predictions of these models are used as features to train a stronger meta-model, which in this research is Gaussian Naïve Bayes. The aim of this process is combining the strengths related to weak models to obtain a more accurate and stable model. The figure 5 illustrates the two-level stacking process:

Level 0: Weak models (LR, KNN, and SVM) are trained on the training data and generate predictions.
Level 1: The predictions of the weak models are utilized as input to train a stronger model (Gaussian Naïve Bayes), which makes the final prediction.

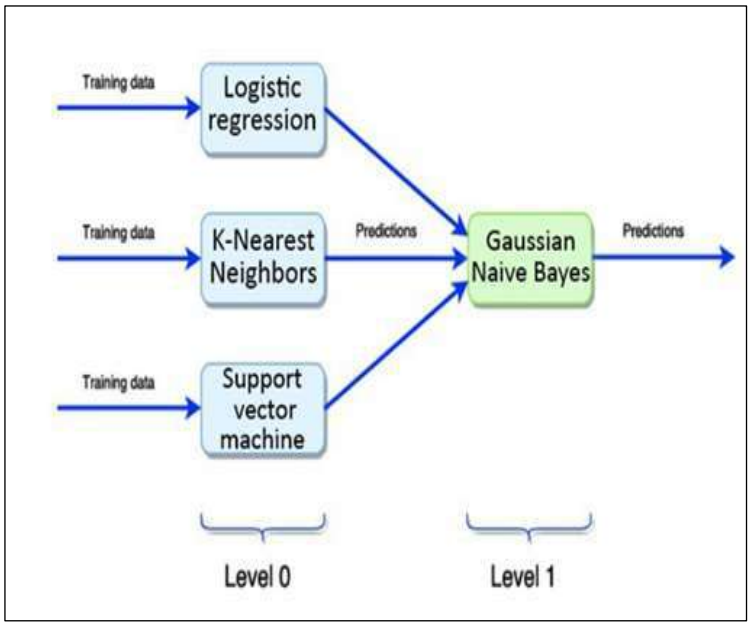


Figure 5: Stacking process.
Source: Authors, (2026).

III.5 ALGORITHM: PSO WITH MODEL (BAGGIN, BOOSTING, STACKING AND VOTING)

Input: Dataset (X, Y), Number of particles (N), Maximum iterations (T), Evaluation criterion (Fitness)

Output: Best feature set, Final model

Step 1. Initialize particles:

- For each particle, randomly initialize a position (feature set) and velocity.

Step 2. Evaluate fitness:

- For each particle, calculate the fitness of the feature set using the evaluation criterion.

Step 3. Update best positions:

- Update the personal best (Pbest) for each particle based on its fitness.

- Update the global best (Gbest) based on the best fitness among all particles.

Step 4. Update velocity and position:

- For each particle, update its velocity using the formula:

Velocity = w * Velocity + c1 * r1 * (Pbest - Position) + c2 * r2 * (Gbest - Position)

- Update the position of the particle based on its new velocity.

Step 5. Repeat steps 2-4:

- Continue for a specified number of iterations (T) or until the stopping criterion is met.

Step 6. Select the best feature set:

- The feature set represented by the global best particle (Gbest) is selected.

Step7. Train Bagging model:

- Use the selected features to train a model on the dataset.

Step 8. Output:

- Return the best feature set and the trained model.

IV. EVALUATION METRICS

In analyzing the results of the proposed models, the use of standard classification metrics was used, like Precision, Accuracy, F1-Score, and Recall. The measures give a detailed insight on the effectiveness of the classifier.

- Accuracy is a general correctness of the model that is specified as the percentage of the correct classification of instances (negative and positive):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

- Precision is the fraction of times positive predictions are made correctly among the instances that are being predicted as positive:

$$precision = \frac{TP}{TP+FP} \tag{3}$$

- Recall (Sensitivity) determines the percentage of correctly recognized positive results relative to the number of all the actual positive cases

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

- F1-Score is the arithmetic mean of Recall and Precision, which provides a balance in the measurement of the two

$$F1 - score = 2 \times \frac{Recall \times precision}{Recall + precision} \tag{5}$$

These measures provide a reasonable comparison between classifiers, particularly when there is imbalance between classes, since they are not only concerned with the overall accuracy but also with the capability of the model to recognize minor classes correctly.

V. RESULTS AND DISCUSSIONS

In the presented section, the experimental results are given as well as analyzed on the basis of the proposed methodology. The data were split into testing and training sets with training set utilized for developing the models and the testing set being used to test the models. The major aim of the evaluation was to look into the efficiency of PSO in the FS alongside ensemble classification methods Table 1 contains the parameters used to evaluate the effectiveness of various optimization techniques in the field of FS through particle fusion, i.e., Bagging, Boosting, Stacking, and Voting. The widely accepted metrics that were used to determine the model performance included Precision, Accuracy, F1-Score, Recall, and the Area Under the Curve (AUC). Furthermore, the results indicate that PSO dramatically enhanced the models through the removal of redundant and irrelevant features hence improving accuracy and inter-computation efficiency.

Boosting outperformed the other ensemble methods by a significant margin, and the generalization was almost perfect with near-perfect results. Bagging and Stacking also demonstrated high results, which testifies to them being reliable ensemble methods. Although the Voting classifier had slightly lower performance compared to the rest, it still had strong and competitive performance. In general, the findings demonstrate the significance of FS in the process of optimization of models. The proposed framework based on the combination of PSO and ensemble techniques proved to be effective and stable in predicting absenteeism. These performances imply the practical situational applicability of the framework in the management of human resources, and organizations can predict the negative effect of employee absenteeism in their productivity as well as operation proactively.

Table 1: PSO Parameters.

Parameter	Description	Value/Default
N	Number of particles	15
T	Maximum number of iterations	20
w	Inertia weight	0.9
c1	Cognitive acceleration factor	2
c2	Social acceleration factor	2
k	Number of neighbors in KNN	7
thres	Threshold for binary conversion	0.5
alpha	Weight for error rate in objective function	0.99
beta	Weight for feature size in objective function	0.01

Source: Authors, (2026).

IV.1 BOOSTING

The Boosting-based model recorded the best performance with an AUC of 1.0 which is a perfect value in terms of classification in figure 6. The convergence of the PSO fitness values over iterations was found to be very rapid, which proves the efficiency of the algorithm in finding the optimal feature subsets. The performance measures were very high with recall (0.987), precision (0.999), accuracy (0.998), and F1-score (0.993) and Boosting was the highest-performing ensemble method used in this study.

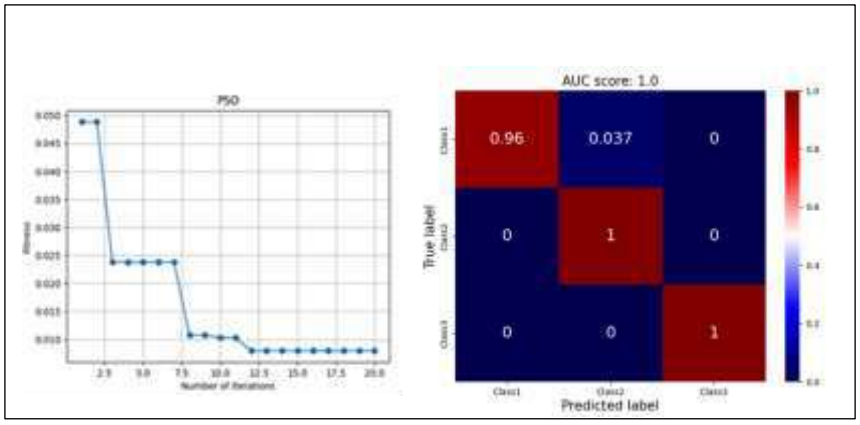


Figure 6: Performance Evaluation of PSO-Based FS and Boosting Classification.
Source: Authors, (2026).

IV.2 BAGGING

The Bagging classifier had an AUC of 0.97 which indicates high classification performance in figure 7. Class 1 had a low level of misclassification (13% to Class 2), whereas Class 2 and Class 3 were classified correctly and 100 per cent. The PSO optimization curve was also characterized by a steady decrease in the values of fitness as the iterations increased maintaining a minimum value, which is indicative of a high degree of convergence. Altogether, Bagging has obtained 0.956 recall, 0.997 precision, 0.993 accuracy, and 0.975 F1-score, which is strong evidence of its strength.

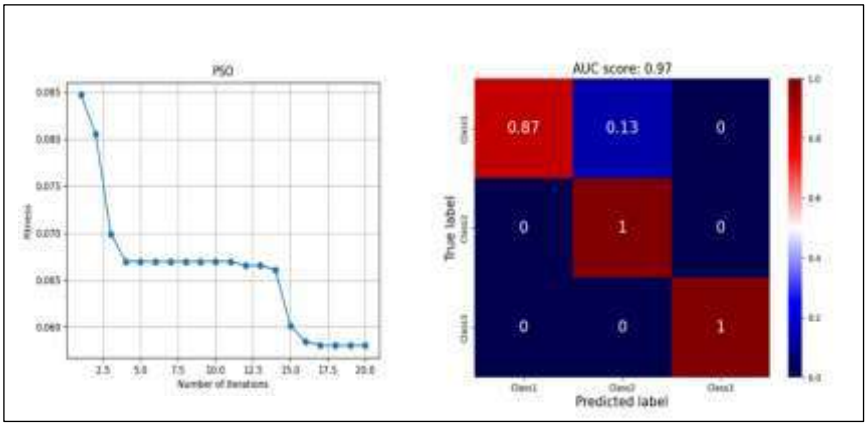


Figure 7: Performance Evaluation of PSO-Based FS and Bagging Classification.
SOURCE: AUTHORS, (2026).

IV.3 STACKING

The Stacking method used weak base models (LR, K-Nearest Neighbors, and SVM), which were in turn followed by a Gaussian Naive Bayes meta-model to fuse their results. The results obtained in this two-level architecture were very reliable with an AUC score of 0.98 in figuer 8 . The technique was able to detect the correct classification of 96 percent of the total data cases which validates that the combination of the multiple learners is effective. The precision, recall, and F1-score of 0.967 as well as the accuracy of 0.990 were the performance metrics used to demonstrate the stability and strength of the stacked model.

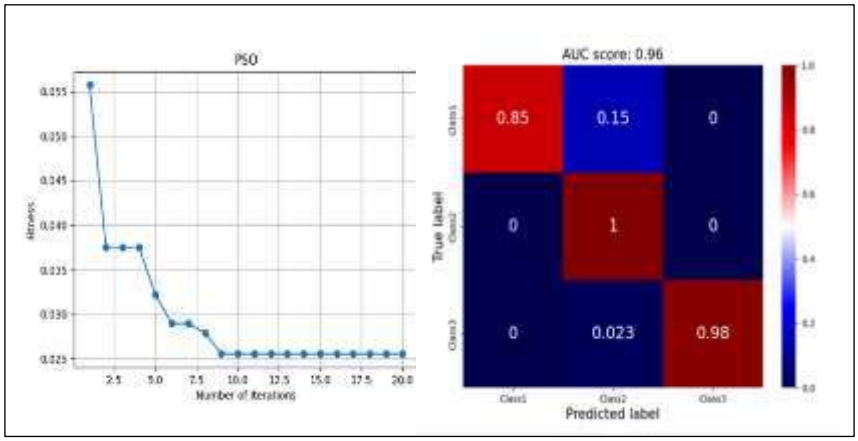


Figure 8: Performance Evaluation of PSO-Based FS and Stacking Classification.
Source: Authors, (2026).

IV.4 VOTING

The Voting classifier used hard voting to combine LR, SVM and NB to come up with the final decision. The approach had an AUC value of 0.981 and was able to classify 98% of the data in figuer 9. Although Voting was less successful in performance than Boosting, it gave precise results with a high recall (0.916), F1-score (0.951), and accuracy (0.982). These findings affirm Voting is a competitive ensemble method as it is used in predicting absenteeism.

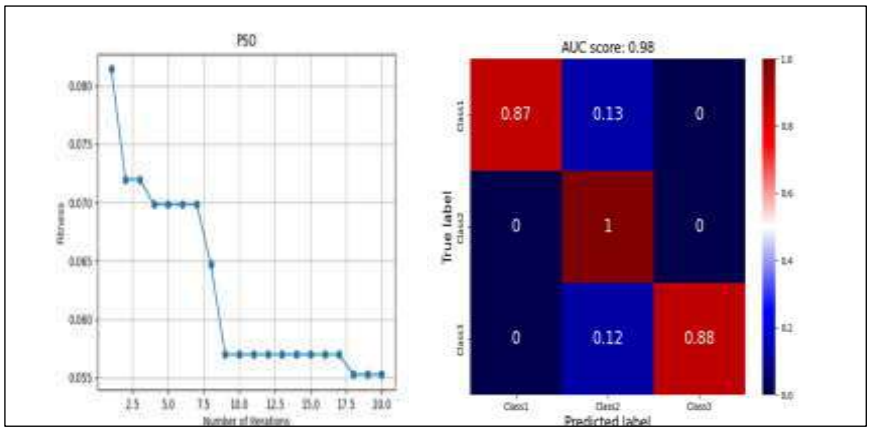


Figure 9: Performance Evaluation of PSO-Based FS and voting.
Source: Authors, (2026).

IV.5 COMPARISON OF MODEL PERFORMANCE METRICS

The performance of the four ensemble techniques Boosting and Bagging Stacking and Voting compares remarkably in table 2. Boosting was the most effective of them with the highest scores in all metrics with recall of 0.987, precision of 0.999, AUC score of 1.0, F1-score of 0.993, and an accuracy of 0.998. Close behind is bagging, which has good results with 0.997 precision, 0.956 recall, 0.975 F1-score, 0.970 AUC score and 0.993 accuracy. Stacking approach was also found to be reliable where the precision is 0.996, recall 0.943, F1-score 0.967, AUC score 0.956 and accuracy is 0.990. Lastly, Voting classifier although slightly lower than the rest, also had competitive scores, showing recall of 0.916, precision of 0.993, AUC score of 0.981, F1-score of 0.951, and its accuracy was 0.982.

These findings affirm that Boosting is the best ensemble technique of predicting absenteeism and Bagging and Stacking have almost similar performance with slight sacrifices in recall. Voting, even though it is slightly lower in performance than the rest, nevertheless, provides good and consistent classification performance.

Table 2: result of evaluation criteria for classification methods.

Method	Precision	recall	f1-score	AUC Score	accuracy
Boosting	0.999	0.987	0.993	1	0.998
Bagging	0.997	0.956	0.975	0.970	0.993
Stacking	0.996	0.943	0.967	0.956	0.990
Voting	0.993	0.916	0.951	0.981	0.982

Source: Authors, (2026).

VI. CONCLUSIONS AND FUTURE WORK

This study proposes a predictive model for employee absenteeism in a courier organization, utilizing PSO for FS combined with ensemble techniques such as bagging, boosting, Stacking, and Voting classification. The methodology involves a three-step workflow: data preprocessing, FS, and ensemble classification using Scikit-learn in Python. Results indicate that PSO effectively removes redundant features and boosts classifier performance, with Boosting achieving the highest accuracy. The model can be used as a decision-support tool by managers and HR specialists in order to control the impacts of absenteeism on productivity through the better management of the workforce. Future studies may include different datasets of various industries, improved FS, incorporation of external variables such as employee satisfaction surveys and use in the real-time decision support systems to improve dynamism in monitoring and predicting absenteeism.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Baydaa Hadi Saoudi, Tabarek Alwan Tuib and Yaqdhan Mahmood Hussein.
Methodology: Baydaa Hadi Saoudi, Tabarek Alwan Tuib and Yaqdhan Mahmood Hussein.
Investigation: Baydaa Hadi Saoudi, Tabarek Alwan Tuib and Yaqdhan Mahmood Hussein.
Discussion of results: Baydaa Hadi Saoudi, Tabarek Alwan Tuib and Yaqdhan Mahmood Hussein.
Writing – Original Draft: Baydaa Hadi Saoudi, Tabarek Alwan Tuib and Yaqdhan Mahmood Hussein.
Writing – Review and Editing: Baydaa Hadi Saoudi, Tabarek Alwan Tuib and Yaqdhan Mahmood Hussein.
Resources: Baydaa Hadi Saoudi, Tabarek Alwan Tuib and Yaqdhan Mahmood Hussein.
Supervision: Baydaa Hadi Saoudi, Tabarek Alwan Tuib and Yaqdhan Mahmood Hussein.
Approval of the final text: Baydaa Hadi Saoudi, Tabarek Alwan Tuib and Yaqdhan Mahmood Hussein.

VIII. REFERENCES

[1] M. A. A. Albadr, A. S. Tiun, N. S. Alseelawi, and R. M. Al-Saedi, "Speech emotion recognition using optimized genetic algorithm-extreme learning machine," Multimedia Tools and Applications, vol. 81, no. 17, pp. 23963–23989, 2022.

[2] A. A. Barreda-Luna, C. J. L. Bustamante, and F. J. Sánchez, "Spatial models and neural networks for identifying sustainable transportation projects: A case study in Querétaro, an intermediate Mexican city," Sustainability, vol. 14, no. 13, p. 7796, 2022.

[3] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," Journal of Intelligent Learning Systems and Applications, vol. 9, no. 1, pp. 1–16, 2017.

[4] T. A. Tuib, B. H. Saoudi, Y. M. Hussein, T. H. Mandeel, and F. T. Al-Dhief, "Convolutional neural network with binary mothflame optimization for emotion detection in electroencephalogram," IAES International Journal of Artificial Intelligence, vol. 13,no. 1, pp. 1172–1178, Mar. 2024, doi: 10.11591/ijai.v13.i1.pp1172-1178.

[5] X. Yang, Z. Ding, and T. Liu, "A survey on deep semi-supervised learning," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 9, pp. 8934–8954, 2022.

[6] S. Dong, P. Wang, and K. Abbas, "A survey on deep learning and its applications," Computer Science Review, vol. 40, p. 100379, 2021.

[7] K. Jha, A. Doshi, P. Patel, and M. Shah, "A comprehensive review on automation in agriculture using artificial intelligence," Artificial Intelligence in Agriculture, vol. 2, pp. 1–12, 2019.

[8] A. Cohen and R. Golan, "Predicting absenteeism and turnover intentions by past absenteeism and work attitudes: An empirical examination of female employees in long term nursing care facilities," Career Development International, vol. 12, no. 5, pp. 416–432, 2007.

[9] P. R. de Jong, Sickness, Disability and Work: Breaking the Barriers—A Synthesis of Findings across OECD Countries. Wiley Online Library, 2011.

[10] M. C. Sturman, Multiple Approaches to Absenteeism Analysis. Ithaca, NY: Cornell University, 1996.

[11] P. Bycio, “Job performance and absenteeism: A review and meta-analysis,” Human Relations, vol. 45, no. 2, pp. 193–220, 1992.

[12] D. A. Harrison and J. J. Martocchio, “Time for absenteeism: A 20-year review of origins, offshoots, and outcomes,” Journal of Management, vol. 24, no. 3, pp. 305–350, 1998.

[13] Maria Maqsood, Saima Anwar Lashari, Murtaja Ali Saare, Sari Ali Sari, Yaqdhan Mahmood Hussein and Hatem Oday Hatem, "Minimization Response Time Task scheduling Algorithm", et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 705 012008, DOI 10.1088/1757-899X/705/1/012008.

[14] C. Murthy, “Employee absenteeism: Its impact on productivity with references to Sukjit Scratch Mills at Nizamabad,” International Journal of Management Research, 2012.

[15] A. M. De Silva, et al., “Feature selection,” in Grammar-Based Feature Generation for Time-Series Prediction, Springer, 2015, pp. 13–24.

[16] A.A. Abdulbari, Z. Zakaria, S.K. Abdul Rahim, Y.M. Hussein, M.M. Jawad, A.M. Hamzah, Design and development broadband monopole antenna for in-doorapplication, TELKOMNIKA 18 (1) (2020) 51, https://doi.org/10.12928/telkomnika.v18i1.13171. Feb.

[17] S. N. Mishra, D. R. Lama, and Y. Pal, “Human resource predictive analytics (HRPA) for HR management in organizations,” International Journal of Scientific & Technology Research, vol. 5, no. 5, pp. 33–35, 2016.

[18] N. Jain and Maitri, “Big data and predictive analytics: A facilitator for talent management,” in Data Science Landscape: Towards Research Standards and Protocols. Springer, 2018, pp. 199–204.

[19] A. Martiniano, et al., “Application of a neuro fuzzy network in prediction of absenteeism at work,” in 7th Iberian Conference on Information Systems and Technologies (CISTI 2012), IEEE, 2012.

[20] N. N. Qomariyah and Y. G. Sucahyo, “Employees’ attendance patterns prediction using classification algorithm: Case study of a private company in Indonesia,” Int’l Journal of Computing, Communications & Instrumentation Engg., vol. 1, no. 1, pp. 1–7, 2014.

[21] T. Gayathri, “Data mining of absentee data to increase productivity,” International Journal of Engineering and Techniques, vol. 4, no. 3, pp. 478–480, 2018.

[22] R. P. Ferreira, et al., “Artificial neural networks and their application in the prediction of absenteeism at work,” International Journal of Recent Scientific Research, vol. 9, no. 1, pp. 2332–2334, 2018.

[23] S. A. Ali Shah, et al., “An enhanced deep neural network for predicting workplace absenteeism,” Complexity, vol. 2020, Article ID 5843932, 2020.

[24] E. L. de Oliveira, et al., “Absenteeism prediction in call centers using machine learning algorithms,” in New Knowledge in Information Systems and Technologies: Volume 1, Springer, 2019, pp. 1–12.

[25] Z. Wahid, et al., “Predicting absenteeism at work using tree-based learners,” in Proc. 3rd Int. Conf. Machine Learning and Soft Computing, 2019.

[26] R. Mythily and D. Mavaluru, “An efficient feature selection algorithm for health care data analysis,” Bulletin of Electrical Engineering and Informatics, vol. 9, no. 3, pp. 877–885, 2020.

[27] R. Kamala and R. J. Thangaiah, “An improved hybrid feature selection method for huge dimensional datasets,” IAES International Journal of Artificial Intelligence, vol. 8, no. 1, pp. 77–85, 2019.

[28] Alaa Abd Ali Hadi, Essam Hamoodi Ahmed, Baydaa Hadi Saoudi, Yaqdhan Mahmood Hussein, Tabarek Alwan Tuib, "Design zigzag edge of S shape slot antenna by using SIW technology for 5G application", TELKOMNIKA Telecommunication Computing Electronics and Control, ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v21i6.23788.

[29] K. Dogruyol and B. Sekeroglu, “Absenteeism prediction: A comparative study using machine learning models,” in Int. Conf. Theory and Application of Soft Computing, Computing with Words and Perceptions, Springer, 2019.

[30] K. Tewari, S. Vandita, and S. Jain, “Predictive analysis of absenteeism in MNCs using machine learning algorithms,” in Proc. ICRIC 2019: Recent Innovations in Computing, Springer, 2020.

[31] A. Rista, J. Ajdari, and X. Zenuni, “Predicting and analyzing absenteeism at workplace using machine learning algorithms,” in 43rd Int. Convention on Information, Communication and Electronic Technology (MIPRO), IEEE, 2020.

[32] M. Skorikov, et al., “Prediction of absenteeism at work using data mining techniques,” in 5th Int. Conf. on Information Technology Research (ICITR), IEEE, 2020.

[33] A. Al-Rasheed, “Identification of important features and data mining classification techniques in predicting employee absenteeism at work,” International Journal of Electrical & Computer Engineering, vol. 11, no. 5, pp. 4083–4092, 2021.

[34] J. Rajath and D. Pandita, “Machine learning as a data science tool to predict absenteeism for factory workers,” in 2022 Int. Conf. on Data Analytics for Business and Industry (ICDABI), IEEE, 2022.

[35] F. Pedregosa, et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1207-1217. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3034

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

META-LEARNING BASED STACKING OF CNN, SVM, AND RANDOM FOREST FOR MULTI-SPECIES CROP LEAF PATHOLOGY RECOGNITION

Ryma Guefrouchi*[1], Siham Amrouch[2]

[1]MISC Laboratory, Constantine 2 University Abdelhamid Mehri Constantine , 25000, Algeria.
[2]LIM Laboratory, Mohamed Cherif Messaadia University Souk Ahras, 41000, Algeria.

[1]https://orcid.org/0000-0002-1756-1256, [2]https://orcid.org/0009-0008-6112-0141

Email: *ryma.guefrouchi@univ-constantine2.dz, s.amrouche@univ-soukahras.dz

ARTICLE INFO

Article History
Received: December 12, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
CNN,
SVM,
RFC,
Ensemble Learning,
Meta-learning,
Crop Disease,
Early Detection.

ABSTRACT

Accurate identification of crop leaf diseases is essential for sustainable agriculture, yet variations in leaf morphology and disease symptoms across species remain challenging for traditional machine learning approaches. This study proposes a meta-learning–based stacking ensemble that integrates Convolutional Neural Networks (CNN), Support Vector Machines (SVM), and Random Forest Classifiers (RFC) for multi-species leaf pathology recognition, evaluated on apple and tomato datasets. Each base model independently learns discriminative visual features, and their class-probability outputs are fused through an RBF-SVM meta-learner. Experimental results show that individual models achieve accuracies of 89% (CNN), 88% (SVM), and 85% (RFC), while the proposed stacking ensemble significantly outperforms them with an accuracy of 98% and macro and weighted F1-scores of 0.975. These results demonstrate the effectiveness of heterogeneous meta-learning ensembles in enhancing robustness and generalization for multi-species crop disease classification.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Crop leaf diseases pose a significant threat to crop yield and quality, making early and accurate detection critical for sustainable agriculture [1], [2]. Traditional diagnosis methods rely on visual inspection by experts, which is time-consuming, labor-intensive, and prone to human error. Automated image-based detection techniques have therefore emerged as promising alternatives, enabling rapid and objective disease monitoring [3]. Over the past decade, deep learning (DL), particularly CNNs, has demonstrated remarkable performance in plant disease recognition. CNNs automatically extract hierarchical visual features such as texture, color variations, and lesion patterns from leaf images, often surpassing traditional handcrafted feature methods [4], [5]. However, purely CNN-based approaches can face challenges, including sensitivity to lighting conditions, background noise, inter-class similarity, and the need for large annotated datasets to avoid overfitting [6]. On the other hand, classical machine learning (ML) models, such as SVMs and RFCs, are known for their robustness on limited datasets and ability to handle high-dimensional feature spaces [7], [8].

Yet, these models typically require handcrafted features or pre-extracted embeddings, limiting their ability to capture complex visual patterns inherent in leaf diseases. To overcome these limitations, hybrid ensemble approaches have gained attention in recent studies. Combining CNN feature extraction with classical classifiers can leverage the strengths of both paradigms: the representational power of deep learning and the robustness and interpretability of classical ML [9], [10]. Moreover, stacking ensembles—where predictions from multiple base learners are fused via a meta-learner—enable capturing complementary information from heterogeneous models, often leading to improved generalization compared to individual models or simple averaging strategies [11]. The primary motivation behind this work is to address the limitations of using a single CNN or classical machine learning model by combining their complementary strengths. The proposed heterogeneous stacking ensemble leverages the hierarchical feature extraction capability of the CNN, the margin-based discriminative power of the SVM, and the robustness of Random Forests.

The outputs of these base models are then fused using an RBF-SVM meta-learner, which captures non-linear relationships among the predicted probabilities and ensures an optimized integration of the individual classifiers' predictions. The approach offers several key

advantages: (i) intelligent weighting of base models per class, allowing the system to exploit the strengths of specific classifiers for different disease categories; (ii) enhanced classification performance, improving accuracy and F1-scores compared to individual models or naive ensembles; and (iii) robustness with limited or imbalanced datasets, ensuring reliable multi-class leaf disease recognition across different species. The remainder of this paper is organized as follows: Section II reviews related work in plant disease detection and ensemble methods. Section III presents the contributions and details of the proposed architectures. Section IV describes the dataset and performance metrics used. Section V presents the experimental results, and Section VI concludes the paper.

II. RELATED WORK

Automated detection of plant leaf diseases has been an active research area due to its potential to reduce manual labor and improve agricultural productivity. Early approaches relied on handcrafted feature extraction methods such as color histograms, shape descriptors, and texture features (GLCM, LBP) combined with classical classifiers like SVM or Random Forest [12], [13]. While these methods are relatively simple and robust on small datasets, they often fail to capture complex patterns of disease lesions, especially under variable environmental conditions.

II.1 DEEP LEARNING FOR PLANT DISEASE RECOGNITION

The emergence of deep learning, particularly Convolutional Neural Networks (CNNs), has significantly advanced plant disease recognition [14]. CNNs automatically learn hierarchical visual features, including texture, color variations, and lesion morphology from leaf images, often outperforming traditional handcrafted feature methods [14], [15]. For instance, the EnConv framework (2024) employs depthwise separable convolutions and inverted residual blocks to achieve 99.87 % accuracy on a dataset of 39 plant disease classes, including apple leaves [16].

Similarly, multi-head CNN architectures (2023) have been developed to enhance robustness and generalization across large and diverse datasets [17]. Despite their success, CNN-based methods still face challenges, such as sensitivity to lighting variations, background noise, inter-class similarity, and the need for large annotated datasets to avoid overfitting [18]. Moreover, most existing deep learning studies focus on single-species datasets, limiting their ability to generalize across multiple crops and disease types.

II.2 HYBRID APPROACHES: CNN + CLASSICAL ML

To address these limitations, recent studies have combined CNN feature extraction with classical machine learning classifiers. For example, an ensemble model integrating CNN feature extractors with Random Forest classifiers (2024) demonstrated improved performance and robustness on heterogeneous or limited datasets [19]. Similarly, a 2025 study proposed a robust ensemble combining multiple deep learning architectures (InceptionResNetV2, MobileNetV2, EfficientNetB3) to exploit their complementary strengths and improve generalization across different datasets [20]. Moreover, hybrid CNN-Vision Transformer frameworks (2025) have been explored to capture both global and local features, reflecting the current trend towards multi-modal and heterogeneous architectures for plant disease detection [21]. Nevertheless, these works typically still focus on a single species, such as apple or tomato, highlighting a need for models capable of handling multiple crop species simultaneously.

II.3 RECENT REVIEWS AND SURVEYS

Recent reviews provide comprehensive summaries of progress and remaining challenges in plant disease detection. A systematic review of deep learning techniques for plant diseases (2024) highlighted both the achievements and challenges, including model generalization to field data and the need for diverse datasets [22]. Another comparative study (2024) evaluated CNN, classical ML, and hybrid methods, showing that ensemble approaches often outperform individual models, particularly under varying environmental conditions [23]. The reviews summarize key studies in plant leaf disease detection, highlighting their methods, targeted crop species, dataset characteristics, and limitations.

Most existing works focus on single-species datasets, which limits their generalizability across multiple crops. While CNN-based approaches achieve high accuracy, they often require large annotated datasets and are sensitive to environmental variations. Hybrid and ensemble methods, such as CNN combined with SVM or Random Forest, improve robustness, but few studies integrate multiple paradigms into a heterogeneous stacking framework capable of handling multi-species disease recognition. This emphasizes the need for our proposed meta-learning based stacking approach.

II.4 RESEARCH GAP

Despite these advances, few studies have simultaneously integrated three paradigms—CNN (as feature extractor and base classifier), classical ML classifiers (SVM and Random Forest on CNN embeddings), and a meta-level fusion via stacking—for crop leaf disease recognition across multiple species. Most existing works focus either on deep architectures (CNN or CNN + ViT) or on homogeneous ensembles for a single crop species, leaving a gap for heterogeneous stacking ensembles that exploit complementary strengths of deep learning and classical machine learning in a multi-species context**.**

III. CONTRIBUTIONS AND DETAILED ARCHITECTURES

Motivated by the critical need for robust and accurate diagnostic tools in precision agriculture, this study introduces a novel heterogeneous stacking ensemble framework for image-based plant disease classification across multiple crop species, utilizing more challenging and diverse datasets to demonstrate the model's robustness and generalizability.The core contribution of this work is the integration of three powerful, yet distinct, learners—CNN, SVM, and RFC—into a cohesive stacking architecture. This integration is strategically designed to capitalize on their complementary advantages:

- CNNs provide superior hierarchical feature extraction from raw image data.
- SVMs offer strong generalization and margin maximization in high-dimensional feature spaces.
- RFCs contribute robust ensemble decision-making and resistance to overfitting.

A meta-learner then synthesizes the predictions from these base models to produce a final, refined classification. Beyond achieving state-of-the-art performance on plant disease datasets, this framework demonstrates applicability across multiple crop species, offering a robust tool for deploying hybrid and ensemble strategies in agricultural image analysis. The following sections describe the architectural components and their interactions, designed for optimal performance in plant disease classification across heterogeneous, multi-species datasets.

III.1 PROPOSED CNN ARCHITECTURE

The custom CNN developed for this study consists of a sequential architecture designed to extract hierarchical visual features and perform robust image classification. The model is divided into two main stages: a feature extraction phase and a classification phase. Figure 1 illustrates the proposed CNN architecture. In the feature extraction stage, the network applies three consecutive convolutional blocks, each using 3×3 kernels and ReLU activation. The first block contains a single Conv2D layer with 32 filters, followed by a 2×2 MaxPooling layer and a Dropout rate of 0.2. The second block increases the representational capacity by stacking two Conv2D layers (with 32 and 64 filters, respectively), followed by a 2×2 MaxPooling layer and 0.25 Dropout. The third block further deepens the architecture with Conv2D layers using 64 and 128 filters, again followed by 2×2 MaxPooling and 0.3 Dropout.

These blocks progressively reduce spatial dimensions while capturing increasingly complex patterns, transitioning from low-level edges and textures to high-level semantic structures. A Flatten layer then transforms the final feature maps into a one-dimensional vector for classification. The classification stage consists of a fully connected Dense layer with ReLU activation, used to learn non-linear combinations of extracted features, followed by a Dropout layer with a higher rate of 0.5 to prevent overfitting at the decision level. The final output layer is a Dense layer with Softmax activation, producing a normalized probability distribution over the target classes. The training settings were configured as follows:

- Optimizer (Adam): chosen for its adaptive learning rate, balancing convergence speed and stability.
- Loss Function (Categorical Cross-Entropy): suitable for multi-class classification.
- Learning Rate: 0.001 the default value used by the Adam optimizer, providing a stable and efficient convergence for most training scenarios.
- Batch Size (32): balances training efficiency and stability.

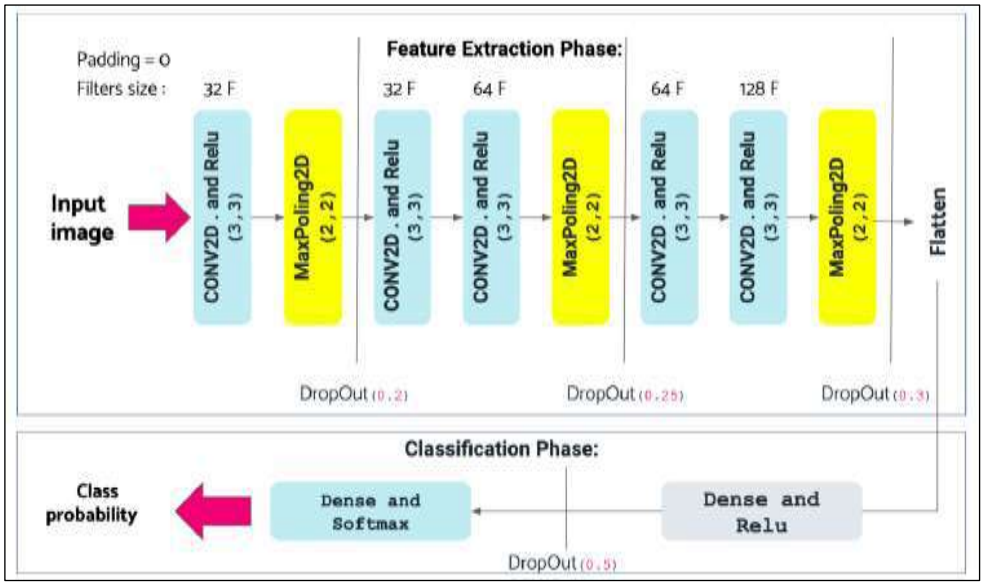


Figure 1: Proposed CNN architecture for crop leaf disease classification.
Source: Authors, (2026).

III.2 SVM MODEL

The SVM model for dataset image classification was developed following a standard preprocessing pipeline. Images were first resized to a uniform dimension and converted to a suitable format for model input. To ensure unbiased evaluation, the dataset was split into training and testing sets. The training set was used to train the SVM model, while the testing set provided an independent basis for performance assessment. Various SVM configurations were explored. Initially, a linear kernel was tested, but it produced suboptimal results. Consequently, alternative kernels were considered, and the Radial Basis Function (RBF) kernel was ultimately adopted, resulting in improved classification accuracy. Further optimization was performed by fine-tuning the regularization parameter, C, to balance model complexity and prevent overfitting. After comprehensive experimentation and evaluation, the SVM configuration yielding the best performance was selected. Figure 2 illustrates the structure and workflow of the finalized SVM model.

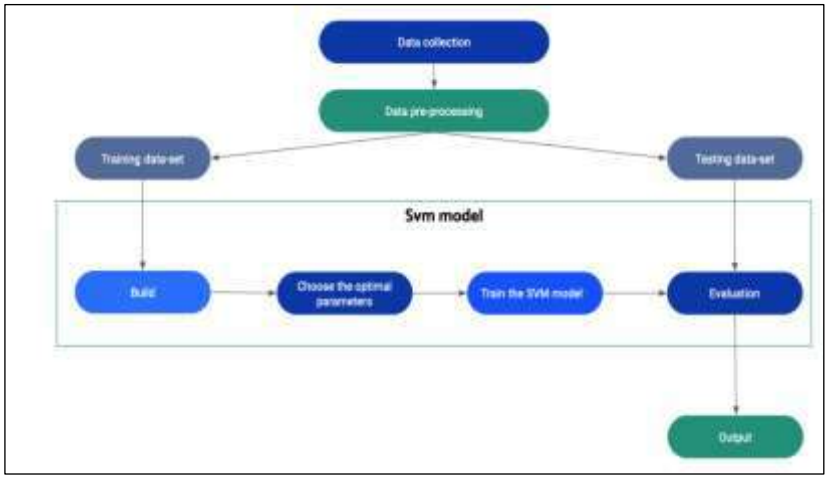


Figure 2: Workflow and structure of the SVM classification model.
Source: Authors, (2026).

III.3 RFC MODEL

The development of the Random Forest Classifier (RFC) model for disease classification began with a standard preprocessing pipeline. All images were resized to a uniform dimension and converted into an appropriate format for model input. Each image was then flattened into a one-dimensional raw pixel vector, which served as the input feature for the RFC. To ensure unbiased evaluation, the dataset was divided into separate training and testing subsets. The training set was used to train the RFC with standard parameters, while the testing set served as an independent dataset for objective performance assessment. This methodology ensured fairness and minimized potential evaluation bias. Figure 3 illustrates the structure of the RFC model.

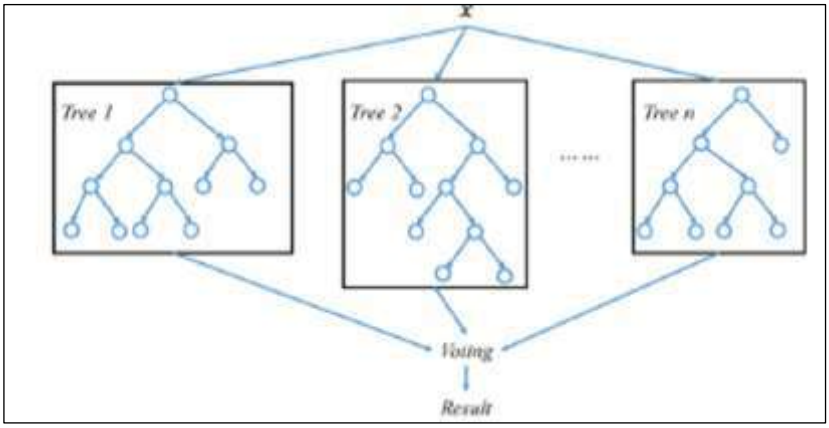


Figure 3: Random Forest classifier architecture.
Source: Authors, (2026).

III.4 ENSEMBLE LEARNING MODEL

The proposed stacking approach introduces an RBF-SVM meta-model which is trained on the outputs of the base classifiers rather than the raw input images. Specifically, for each training sample, the class probability predictions from CNN, SVM, and RFC are concatenated to form a meta-feature vector. The SVM with a RBF kernel is then fitted on these meta-features and the corresponding true class labels. The RBF kernel enables the SVM to capture non-linear relationships among the base classifier outputs. To achieve optimal performance, the hyperparameters C (regularization parameter) and γ (RBF kernel coefficient) are tuned using cross-validation. This procedure ensures that the meta-model generalizes well, effectively integrating the complementary strengths of the base classifiers to improve the overall prediction accuracy.Figure 4 illustrates the proposed meta-learning based architecture and Algorithm 1 details different steps of the proposed.

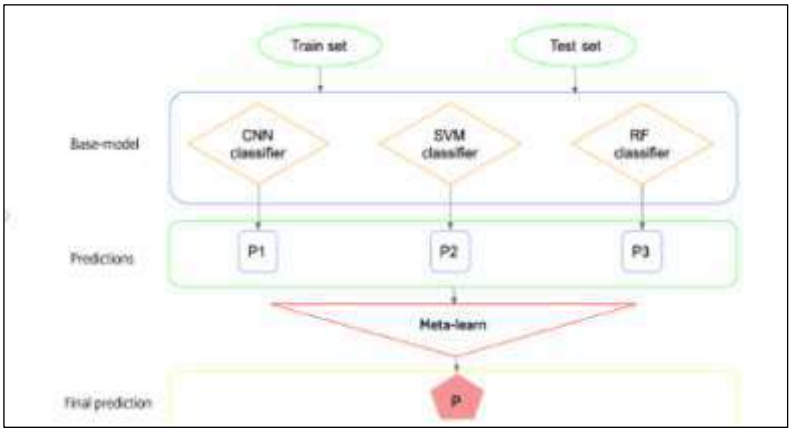


Figure 4: RBF-SVM Meta-learning architecture.
Source: Authors, (2026).

Algorithm 1: Stacking Ensemble with RBF-SVM Meta-Model.

Input: Training dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$
Base classifiers: CNN, SVM, RFC
Meta-model: SVM with RBF kernel

Output: Trained stacking ensemble model

Steps:

1. Train Base Classifiers:
   - Train CNN, SVM, and RFC on D.
2. Generate Base Predictions:
   - For each $x_i \in$ D, compute probabilities: $P_i^{CNN}, P_i^{SVM}, P_i^{RFC}$
3. Prepare Meta-Model Input:
   - Concatenate probabilities: $P_i = [P_i^{CNN}, P_i^{SVM}, P_i^{RFC}]$
4. Train Meta-Model:
   - Fit RBF-SVM on $\{(x_i, y_i)\}$
   - Optimize C and γ via cross-validation
5. Prediction:
   - For new sample x_{new}, compute base probabilities
   - From $P_{new} = [P_{new}^{CNN}, P_{new}^{SVM}, P_{new}^{RFC}]$, predict final class: $\widehat{Y_{new}}) = SVM_{RBF}(P_{new})$
   - End

Source: Authors, (2026).

IV. DATASET AND METRIC PERFORMANCE

IV.1 DATA COLLECTION

The dataset used in this study comprises images of apple and tomato leaves sourced from the publicly available PlantVillage repository [24]. All images are fully anonymized, and their collection complies with open-access ethical guidelines.
The dataset is structured as follows:

- Apple Dataset: Comprising 9,714 images of apple leaf disease .
- Tomato Dataset: Containing 20,661 images across ten tomato leaf disease classes.

Figure 5 shows the sample images representing ten classes of tomato leaf disease, and Figure 6 shows the four cases of apple disease. Tables 1 and 2 summarize the image distributions for the tomato and apple datasets, respectively.

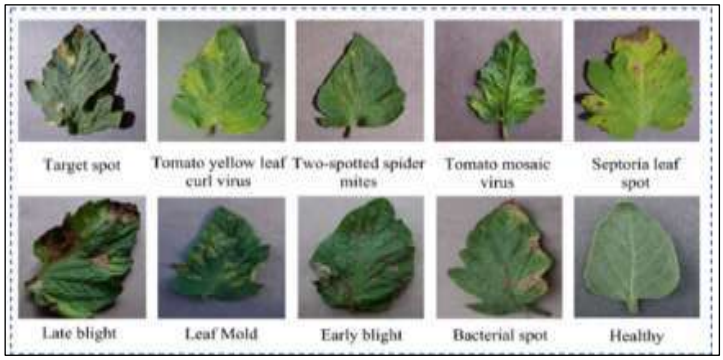


Figure 5: Tomato leaf disease classes used in the dataset.
Source: Authors, (2026).

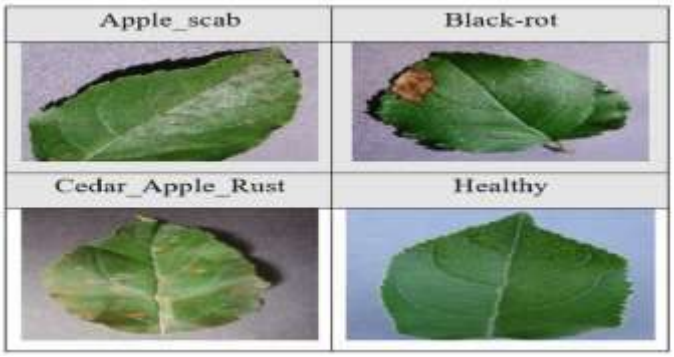


Figure 6: Apple leaf disease classes used in the dataset.
Source: Authors, (2026).

Table 1: Distribution of Images across the Disease Categories.

Disease Category	Number of Images
Early Blight	1136
Late Blight	2168
Septoria Leaf Spot	2013
Bacterial Spot	2431
Leaf Mold	1071
Yellow Leaf Curl Virus	6137
Mosaic Virus	405
Target Spot	1591
Two-spotted-spider mite	1904
Healthy	1805

Source: Authors, (2026).

Table 2: Distribution of apple images across the disease categories.

Diseases	Number of images
Apple Scab	2520
Black Rot	2484
Cedar Apple Rust	2200
Healthy	2510

Source: Authors, (2026).

IV.2 DATA SPLITTING AND DATA PREPROCESSING

Each dataset was split into three subsets: 70 % for training, 15 % for validation, and 15 % for testing. All images were resized to 64 × 64 pixels and normalized to the [0, 1] range by scaling pixel intensities by 1/255. To improve model generalization and mitigate overfitting, on-the-fly data augmentation was applied to the training subset. The augmentation parameters were carefully defined for reproducibility and included: rotation (±20°), zoom scaling (0.8–1.2), shear deformation (0.1), horizontal flipping, and brightness adjustments (±15%). The validation and test sets were only normalized, without any augmentation, to ensure unbiased evaluation. No external images were added to these subsets, preserving the integrity of the experimental setup. Additionally, to address class imbalance, particularly within the tomato dataset, a class-weighted loss function was employed during training. This ensures that both minority and majority classes contribute proportionally to the optimization process, enhancing overall classification robustness.

IV.3 PERFORMANCE METRICS

To evaluate our models, we used the well-known metrics derived from the confusion matrix, based on true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The following standard performance metrics were computed:

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$Accuracy = \frac{TP+TN}{Total} \tag{3}$$

$$F1-score = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{4}$$

V. RESULTS AND DISCUSSIONS

V.1 EXPERIMENT RESULTS

In this section of our thesis, our primary objective is to substantiate the effectiveness of our multi-species disease classification system, which leverages ensemble learning techniques. We aim to present empirical evidence supporting the validity and performance of our proposed approach. Additionally, we provide a comprehensive overview of the development tools and techniques employed during the implementation of our solution. Table 3 summarizes the hardware specifications of the personal computer (PC) used to conduct all experiments.

Table 3: PC's characteristics.

Processor	AMD Ryzen 9 7940HS
RAM	16 GB
Graphics Card	NVIDIA GeForce RTX 2070
Storage	15 GB

Source: Authors, (2026).

- CNN model

The CNN model demonstrated the strongest performance among the three base models, achieving an overall accuracy of 89%. Both the macro and weighted averages for precision, recall, and F1-score consistently fall within the 0.887–0.889 range (Table 5). This indicates robust and balanced performance across all classes. The marginally higher weighted averages suggest the model is particularly effective at classifying the more populous classes, such as Yellow Leaf Curl Virus and Tomato-Healthy (Table 4). These results underscore the CNN's superior capability in extracting complex hierarchical features from image data.

Table 4: Classification Performance per Class for CNN model.

	Class	**Precision**	**Recall**	**F1-score**
Apple Classes	Apple-Scab	0.86	0.82	0.84
	Black-Rot	0.86	0.89	0.88
	Cedar-Apple-Rust	0.94	0.94	0.94
	Apple-Healthy	0.89	0.90	0.89
Tomato Classes	Early Blight	0.88	0.86	0.87
	Late Blight	0.89	0.88	0.88
	Septoria Leaf Spot	0.90	0.88	0.89
	Bacterial Spot	0.87	0.86	0.86
	Leaf Mold	0.88	0.87	0.87
	Yellow Leaf Curl Virus	0.91	0.90	0.90
	Mosaic Virus	0.89	0.88	0.88
	Target Spot	0.87	0.86	0.86
	Two-Spotted Spider Mite	0.88	0.87	0.87
	Tomato-Healthy	0.90	0.89	0.89

Source: Authors, (2026).

Table 5: Aggregated performance metrics CNN model.

Metric	Precision	Recall	F1-score
Accuracy	—	—	89%
Macro Avg	0.8875	0.8875	0.8875
Weighted Avg	0.887	0.889	0.889

Source: Authors, (2026).

- SVM model

The SVM model achieved solid performance, with an overall accuracy of **88%** (Table 7)**.** Its macro and weighted averages for all metrics are closely aligned near 0.867–0.869, indicating consistent performance without significant bias towards any particular class (Table 6). While its metrics are consistently approximately 0.02 lower than the CNN's across the board, it remains a strong and reliable model for this classification task.

Table 6: Classification Performance per Class for SVM model.

	Class	**Precision**	**Recall**	**F1-score**
Apple Classes	Apple-Scab	0.84	0.80	0.82
	Black-Rot	0.85	0.87	0.86
	Cedar-Apple-Rust	0.92	0.92	0.92
	Apple-Healthy	0.87	0.88	0.87
	Early Blight	0.86	0.84	0.85

Tomato Classes	Late Blight	0.87	0.86	0.86
	Septoria Leaf Spot	0.88	0.86	0.87
	Bacterial Spot	0.85	0.84	0.85
	Leaf Mold	0.86	0.85	0.85
	Yellow Leaf Curl Virus	0.89	0.88	0.88
	Mosaic Virus	0.87	0.85	0.86
	Target Spot	0.85	0.84	0.84
	Two-Spotted Spider Mite	0.86	0.85	0.85
	Tomato-Healthy	0.88	0.87	0.87

Source: Authors, (2026).

Table 7: Aggregated performance metrics SVM model.

Metric	Precision	Recall	F1-score
Accuracy	—	—	88%
Macro Avg	0.869	0.867	0.867
Weighted Avg	0.868	0.868	0.868

Source: Authors, (2026).

- RFC model

The RFC model achieved a baseline accuracy of **85%**. Its performance metrics macro and weighted averages ~0.83 (Table 9) are the lowest among the three individual models. The per-class F1-scores show a more pronounced drop, particularly for classes like Apple-Scab (0.78) and Target Spot (0.79) (Table 8), indicating areas where the model struggles compared to the CNN and SVM.

Table 8: Classification Performance per Class for RFC model.

	Class	Precision	Recall	F1-score
Apple Classes	Apple-Scab	0.80	0.76	0.78
	Black-Rot	0.82	0.83	0.82
	Cedar-Apple-Rust	0.88	0.87	0.87
	Apple-Healthy	0.84	0.85	0.84
Tomato Classes	Early Blight	0.82	0.80	0.81
	Late Blight	0.84	0.83	0.83
	Septoria Leaf Spot	0.85	0.82	0.83
	Bacterial Spot	0.81	0.80	0.80
	Leaf Mold	0.82	0.80	0.81
	Yellow Leaf Curl Virus	0.86	0.85	0.85
	Mosaic Virus	0.83	0.81	0.82
	Target Spot	0.80	0.79	0.79
	Two-Spotted Spider Mite	0.82	0.80	0.81
	Tomato-Healthy	0.85	0.84	0.84

Source: Authors, (2026).

Table 9: Aggregated performance metrics RFC model.

Metric	Precision	Recall	F1-score
Accuracy	—	—	85%
Macro Avg	0.833	0.828	0.828
Weighted Avg	0.833	0.832	0.832

Source: Authors, (2026).

- RBF-SVM meta-learning model

As showing in Tables 10 and 11, the proposed ensemble learning achieved remarkably high performance, with macro and weighted averages above 0.975 across all metrics, indicating both high precision and strong generalization.

Table 10: Classification Performance per Class for RBF-SVM meta-learning.

Class Type	Class	Precision	Recall	F1-score
Apple Classes	Apple-Scab	0.97	0.95	0.96
	Black-Rot	0.99	0.98	0.99
	Cedar-Apple-Rust	0.97	0.99	0.98
	Apple-Healthy	0.97	0.97	0.97
	Early Blight	0.98	0.97	0.97

Tomato Classes	Late Blight	0.98	0.97	0.97
	Septoria Leaf Spot	0.97	0.97	0.97
	Bacterial Spot	0.96	0.95	0.96
	Leaf Mold	0.97	0.96	0.96
	Yellow Leaf Curl Virus	0.98	0.97	0.97
	Mosaic Virus	0.97	0.96	0.97
	Target Spot	0.96	0.95	0.96
	Two-Spotted Spider Mite	0.97	0.96	0.96
	Tomato-Healthy	0.98	0.97	0.97

Source: Authors, (2026).

Table 11: Aggregated performance metrics RBF-SVM meta-learning model.

Metric	Precision	Recall	F1-score
Accuracy	—	—	98%
Macro Avg	0.975	0.9725	0.975
Weighted Avg	0.975	0.9725	0.975

Source: Authors, (2026).

The ensemble model significantly outperforms all individual models. By leveraging complementary strengths—CNN's feature extraction, SVM's margin-based generalization, and RFC's ensemble diversity—the stacking approach improves both minority and majority class predictions, achieving near-perfect performance. This demonstrates the effectiveness of ensemble learning in multi-class leaf disease classification. As illustrated in Figure 7, the accuracy curves show a consistent upward trend for both the training and validation sets across the epochs.

The validation curve closely follows the training curve, suggesting that the model learns meaningful representations without exhibiting overfitting. After a rapid improvement during the initial epochs, both curves stabilize and converge toward values exceeding 98%. Similarly, the loss curves display a steady decrease for both training and validation data. Throughout most epochs, the validation loss curve remains slightly below the training loss curve, indicating stable learning dynamics and good generalization capability. The parallel behavior of the two curves and the absence of divergence further confirm that the model maintains reliable performance without signs of overfitting.

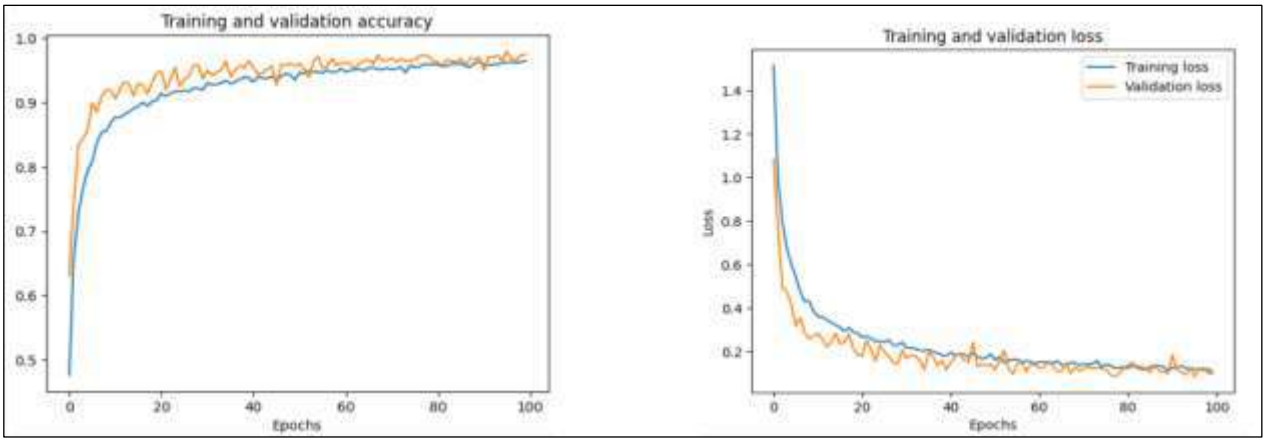


Figure 7: Training and validation accuracy and loss curves for RBF-SVM Meta-learning model.
Source: Authors, (2026).

V.2 DISCUSSION AND COMPARISON

In this study, an RBF-SVM was employed as the meta-learner to integrate predictions from three independently trained base models—CNN, SVM, and RF—across multi-crop leaf disease datasets. Individually, the CNN, SVM, and RF achieved accuracies of 89%, 88%, and 85%, respectively. Each base model contributed complementary information: the CNN extracted hierarchical visual features, while the SVM and RF leveraged raw pixel vectors for independent predictions. By synthesizing the predictions from the individual base models, the RBF-SVM meta-learner achieved a significantly higher overall accuracy of 98%, as illustrated in Figure 8, effectively handling more challenging and diverse datasets, including multiple crop species with varying leaf shapes, disease patterns, and imaging conditions. This stacking ensemble strategy demonstrates both robustness and generalizability, highlighting the effectiveness of meta-learning for complex, multi-species disease recognition tasks.

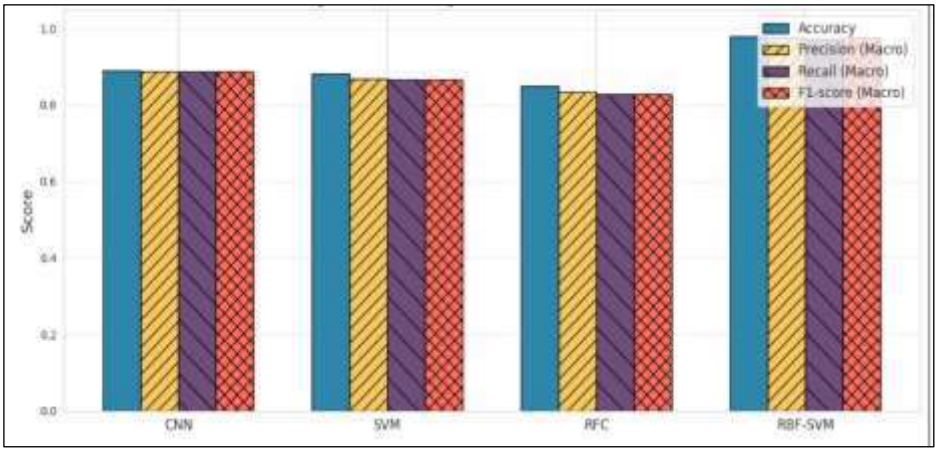


Figure 8: Performance comparison between the base models and RBF-SVM meta-learner model.
Source: Authors, (2026).

Table 12 compares our results with state-of-the-art approaches evaluated on the same dataset (PlantVillage). The table demonstrates that our proposed heterogeneous stacking ensemble outperforms or matches the performance of recent state-of-the-art methods while addressing key limitations. Previous studies [20–23] primarily focus on CNN-based architectures, hybrid CNN-classical ML models, or CNN + Vision Transformer approaches. These methods achieve high reported accuracies, typically in the 95–99% range, with Liang et al. [21] reporting 99.24% on apple leaves and ~98% on corn. However, these studies are largely restricted to single-species or controlled datasets, limiting their generalization to diverse, real-world conditions.

In contrast, our approach combines CNN, SVM, and Random Forest classifiers through an RBF-SVM meta-learner, achieving 98% overall accuracy across multiple crop species, including apple and tomato. By leveraging the complementary strengths of deep learning and classical ML models, the ensemble is able to maintain high accuracy on more challenging, multi-species datasets, demonstrating robustness and generalization beyond conventional single-model approaches. This comparison highlights that, while previous methods perform well on controlled datasets, our stacking ensemble provides a versatile and reliable framework for multi-species plant disease classification, making it well-suited for practical applications in precision agriculture.

Table 12: Comparing our ensemble with related works.

Ref	Year	Approche / Models	Dataset / Species	Reported Accuracy
[20] A. Kumar & D. Singh	2024	Comparative study: CNN vs classical ML vs hybrid	Various crops	CNN ~97%, Hybrid ~91.7%
[21] X. Liang et al.	2025	Hybrid CNN + Vision Transformer	Apple, Corn	Apple 99.24%, Corn ~98%
[22] S. Sharma & V. Singh	2023	Deep learning: CNN & hybrid approaches	Various crops	95–99% (varies by study)
[23] R. Gupta & S. Kumar	2024	CNN, classical ML, hybrid	Leaf disease datasets	95–99%
RBF-SVM meta-learning	2025	CNN, SVM, RFC	Apple, Tomato	98% Multi-species

Source: Authors, (2026).

VI. CONCLUSIONS

This study proposes a meta-learning–based stacking ensemble for multi-species crop leaf pathology recognition, integrating Convolutional Neural Networks (CNN), Support Vector Machines (SVM), and Random Forest Classifiers (RFC) as heterogeneous base learners. Each model was trained independently using its own feature representation: the CNN learned hierarchical visual features directly from images, while the SVM and RFC were trained on raw pixel vectors obtained from resized and flattened images. Their individual predictions were then combined by a meta-level SVM, which learned to exploit complementary strengths across the three models. This stacking strategy resulted in a significant improvement in overall classification performance compared to any single model alone.

Experimental results demonstrated that while individual models achieved accuracies between 85 % and 89 %, the stacking ensemble reached near-perfect performance with macro and weighted F1-scores of 0.975, highlighting its ability to generalize across multiple crop species, including apple and tomato. These results confirm that combining diverse learning paradigms via meta-learning effectively addresses the limitations of single-model approaches, particularly in multi-class and multi-species scenarios. The proposed framework provides a robust, accurate, and generalizable tool for automated crop disease detection, which can support precision agriculture and early disease management. Future work will focus on extending this approach to additional crop species, incorporating field-acquired images under variable conditions, and exploring lightweight architectures to enable real-time deployment on mobile or edge devices.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Ryma Guefrouchi and Siham Amrouch.
Methodology: Ryma Guefrouchi and Siham Amrouch.
Investigation: Ryma Guefrouchi and Siham Amrouch.
Discussion of results: Ryma Guefrouchi and Siham Amrouch.
Writing – Original Draft: Ryma Guefrouchi and Siham Amrouch.
Writing – Review and Editing: Ryma Guefrouchi and Siham Amrouch.
Resources: Ryma Guefrouchi and Siham Amrouch.
Supervision: Ryma Guefrouchi and Siham Amrouch
Approval of the final text: Ryma Guefrouchi and Siham Amrouch.

VIII. REFERENCES

[1] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," Comput. Electron. Agric., vol. 145, pp. 311–318, 2018.

[2] S. Sladojevic et al., "Deep neural networks based recognition of plant diseases by leaf image classification," Comput. Intell. Neurosci., Art. no. 3289801, 2016.

[3] P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," Front. Plant Sci., vol. 7, p. 1419, 2016.

[4] E. C. Too et al., "A comparative study of fine-tuning deep learning models for plant disease identification," Comput. Electron. Agric., vol. 161, pp. 272–279, 2019.

[5] S. Brahimi, D. Boukhalfa, and A. Moussaoui, "Deep learning for tomato diseases: Classification and symptoms visualization," Appl. Artif. Intell., vol. 31, no. 4, pp. 299–315, 2017.

[6] A. Ramcharan et al., "Deep learning for image-based cassava disease detection," Front. Plant Sci., vol. 8, p. 1852, 2017.

[7] V. Singh and R. K. Sharma, "Plant leaf disease detection using feature extraction and machine learning techniques: A review," Int. J. Adv. Res. Comput. Sci., vol. 8, no. 5, pp. 78–84, 2017.

[8] L. Zhang, Y. Yin, and F. Wang, "Recognition of plant leaf diseases based on support vector machine and convolutional neural network," Int. J. Agric. Biol. Eng., vol. 12, no. 2, pp. 76–84, 2019.

[9] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," Comput. Electron. Agric., vol. 147, pp. 70–90, 2018.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015. doi: 10.1038/nature14539.

[11] S. Wang, J. Zhang, and Y. Liu, "Ensemble deep learning: A review," Neurocomputing, vol. 439, pp. 54–68, 2021.

[12] A. Sharma and R. C. Poonia, "Recent advances in automated plant leaf disease detection using machine learning," Mater. Today: Proc., vol. 56, pp. 1234–1240, 2022.

[13] V. Singh and R. Kaur, "A review of machine learning and deep learning methods for plant disease detection," Comput. Electron. Agric., vol. 200, p. 107233, 2022.

[14] J, J., & R, B. (2021). Detection of plant leaf diseases using recent progress in Deep Learning-Based identification techniques. ITEGAM-JETIA, 7(30), 29-36. https://doi.org/10.5935/jetia.v7i30.768
H. Li, J. Wang, and X. Chen, "EnConv: Enhanced CNN framework for plant leaf disease classification," Comput. Electron. Agric., vol. 210, p. 107850, 2024.

[15] Y. Zhang, L. Xu, and H. Zhao, "Multi-head CNN for robust plant disease detection," Neural Comput. Appl., vol. 35, pp. 10875–10890, 2023.

[16] S. Patel and P. Kumar, "Ensemble model for grape leaf disease detection using CNN feature extractors and random forest," Comput. Electron. Agric., vol. 210, p. 107805, 2024.

[17] X. Zhou, M. Li, and F. Yang, "A robust ensemble model for plant disease detection using deep learning architectures," Agriculture, vol. 15, p. 159, 2025.

[18] P. Li, J. Chen, and L. Huang, "A hybrid framework for plant leaf disease detection using CNN and vision transformer," Artif. Intell. Rev., vol. 58, pp. 4123–4140, 2025.

[19] Q. Wang, Y. Zhang, and S. Liu, "Systematic review of deep learning techniques for plant disease detection," Appl. Sci., vol. 14, p. 4512, 2024.

[20] A. Kumar and D. Singh, "Comparative study of CNN, classical ML, and hybrid methods for plant disease classification," J. Big Data, vol. 11, p. 98, 2024.

[21] X. Liang, Y. Chen, and W. Zhao, "Hybrid CNN-vision transformer model for plant leaf disease detection," J. Intell. Fuzzy Syst., vol. 48, pp. 4123–4135, 2025.

[22] S. Sharma and V. Singh, "Plant Disease Detection Using Deep Learning: Recent Trends and Challenges," Computers and Electronics in Agriculture, vol. 208, p. 107850, 2023.

[23] R. Gupta and S. Kumar, "Performance Comparison of CNN, Machine Learning, and Hybrid Methods for Leaf Disease Classification," Neural Computing and Applications, vol. 36, pp. 15021–15038, 2024.

[24] PlantVillage Dataset." [Online]. Available: https://www.kaggle.com/datasets/emmarex/plantdisease.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1218-1228. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3142

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

THE FOUNDATION OF A DEFECT TRACKING MATRIX AND CORRECTIVE MEASURES FOR DEFECT REDUCTION IN THE CRANKSHAFT FORGING INDUSTRY: A PATH TOWARD SUSTAINABLE MANUFACTURING PRACTICES & DRIVING INNOVATION IN QUALITY CONTROL

Amitkumar B. Solanki[*1], Nisha Bora[2], Nilesh H. Pancholi[3] and Nilesh R. Tank[4]

[1, 4]Mechanical Engineering Department , Government Engineering College, Bhavnagar, Gujarat, INDIA
[2, 3]Mechanical Engineering Department, L. D. College of Engineering, Ahmedabad, Gujarat, INDIA.

[1]https://orcid.org/0000-0002-9723-1766, [2]https://orcid.org/0009-0007-3619-6462
[3]http://orcid.org/0000-0003-3420-0724, [4]http://orcid.org/0009-0002-8273-0527

Email: *solankiamit4239@gmail.com

ARTICLE INFO

Article History
Received: December 15, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Forging Industry,
Quality Control,
Defect Tracking Matrix,
Innovation,
Crankshaft

ABSTRACT

Throughout the first decade of this century, the forging industry has experienced rapid expansion, with an increased demand for high-quality forged products requiring zero defects. This article lays the groundwork for limiting forging defects in the crankshaft forging industry sector. Success in this field demands a small rejection or scrap percentage, shortened lead times, and minimal waste in part production. Using a defect tracking matrix, this study identifies common defects such as underfills, laps, and cracks, which are primarily caused by improper material handling, inadequate die design, and insufficient process control. This defect tracking matrix, or cause effect matrix, has the ability to serve as a profound aid for both the design engineer and the producer for any and all aspects of forging and work piece topography that are destined for use in industry sectors that deal with crankshaft production. Key strategies to address these issues include optimizing die design, enhancing process monitoring, and improving operator training. The implementation of corrective measures led to a significant reduction in defects, with the final rejection rate decreasing from 4.57% to 1.93%. The most notable reductions were in defects like underfills, laps, and cracks. This approach not only improves product quality but also enhances overall efficiency in crankshaft manufacturing. The methodologies and strategies outlined can be broadly applied across the forging industry to achieve similar improvements in defect reduction and process performance.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

The forging industry has expanded considerably during the course of this century. Most of the credit for this growth goes to the rapid emergence of more specialized technology, better materials, and heftier presses. Because of this, conflict among forging enterprises has heated up, and 'precise 1st time' products have become a top priority. Using forging systems quickly to make more assemblies has increased the demand for forgings that are free of flaws [1]. The method of metal forging is an effective one for the production of by facilitating a robust structural parameter that meets compliance with the channel flow. Mass manufacturing of accurate components, including crankshafts and connecting rods, may be accomplished at a reduced cost. Forging deforms metals plastically using regionalized compressive forces. Manual or power hammers, presses, or other forging-specific machines apply these forces. Forging transforms a billet or ingot into the desired product [2]. When defects show up during the forging process, they need to be fixed quickly in order to meet both industry standards and the needs of local customers. To satisfy the requirements of the industrial evaluation criteria and the service standard, forging shortcomings are required to be resolved quickly.

During all stages of the forging process, this sector has a lot of problems, such as mis-location, laps, mismatch, un-filling, and seam cracks [3]. Getting closer to this goal could be helped by having a reliable model aid system that helps people avoid the most common kinds of mistakes. Because it's hard for a designer to find the best way to shape a product in as few steps as possible [4], this technology needs to be made. The metal handbook [5] specifies the following steps for forging procedures to be performed: trimming forging parts to remove flashes from the final forged part; hammering or pressing the hot billet in various imprint dies. The quality of a forged product depends on the billet's metal and configuration, die and material temperature, and the quantity and structure of intermediary dies.

Forged crankshafts are employed when load stress is applied, and mitigating measures are vital [6]. Within an internal combustion engine, there is a significant part called the crankshaft that rotates at a rapid speed [7]. The crankshaft is a crucial component of diesel engines and cars, and its complicated design includes cylinders that serve as bearings and arms that serve as crank webs, and it has a direct role in determining the engine's overall performance in terms of both power and efficiency [8]. It turns the reciprocating motion of the piston into rotary motion and then back into reciprocating motion with a simple mechanism. [9] The primary processes involved in producing the crankshaft in the forging industry are shown in (Figure 1). In this process, after heating, the billets are placed beneath the upper and lowest dies and pressed through into the desired form.

This happens in three steps: first, in the roller section, where the part is pre-shaped; second, in the blocker, where the part is rough-forged; and third, in the finisher die, where the finished shape is done. Based on analytical and practical designs as well as real experience, this article discusses the basic necessity of a defect tracking system. While defects may be found using simulations, their origin is still mostly unknown. Because of this, it would be helpful to have a system that lists problems, what makes them happen, and how to fix them. Forging defects and their likely sources are catalogued and organised into categories with accompanying guidance in the cause-effect matrix or by using a defect tracking system. This Defect-Tracking (Cause Effect) Matrix might be implemented into an intelligent system that directs the factory towards achieving "valid 1st time" output.

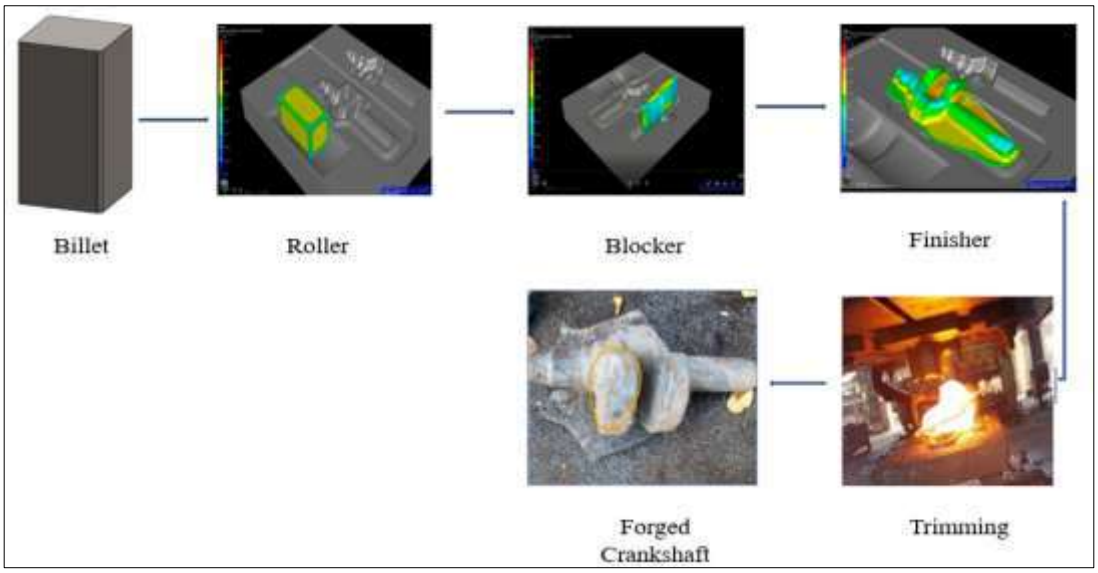


Figure 1: Vital crankshaft forging stages.
Source: Authors, (2026).

II. MATERIALS AND METHODS

II.1 MATERIALS FOR THE FORGED STEEL CRANKSHAFT

Materials for the crankshaft should be readily formed and produced, in addition to having sufficient strength, durability, and rigidity, as well as fatigue resistance power. Ingots of metal(steel) are formed into predetermined forms using specialized presses and other forging equipment to produce crank shafts. Because of advances in the technology used to make steel, it is now feasible to eliminate contaminants in the materials used to create crankshafts [10], [11-12]. Crankshafts are often produced from AISI 1040, AISI 1045, or a comparable kind of steel that is forged [13-14]. Medium-carbon steel alloys contain iron and 0.25-0.45 percent carbon, as well as chromium, nickel, manganese, silicon, cobalt, molybdenum, vanadium, and other elements. Manganese-molybdenum steel is an inexpensive forging steel used for petrol engine crankshafts; chromium-molybdenum steel for light commercial diesel engines; and nickel-chromium-molybdenum steel for heavy-duty diesel engines. Nickel-chrome-moly alloy AISI 4340 is a popular choice for high-performance crankshafts owing to its high strength, fatigue resistance, and ductility [15-17].

Crankshafts are typically made from various types of forged steel grades. Figure 2 illustrates a comparison of mechanical properties among different crankshaft materials. The development of new materials aims to reduce costs by utilizing enhanced micro-alloyed steels. These steels eliminate the need for heat treatment and offer improved material qualities, providing more options for product design. The assessed micro-alloy grades show potential cost savings compared to conventional materials such as tempered and quenched alloy steel (AISI 4140) or carbon steel (AISI 1050), with a reduction of 11% to 19%.In terms of performance, the fatigue strength of micro-alloy grades is generally equal to or higher than that of existing materials, while maintaining superior machinability. This makes micro-alloyed steel a common substitute for forged steel. By using 38MnVS6 or VANARD 925 air-cooled micro-alloyed steel instead of the typical AISI 4142 steel, the cost of the finished crankshaft can be reduced by 13% [18].

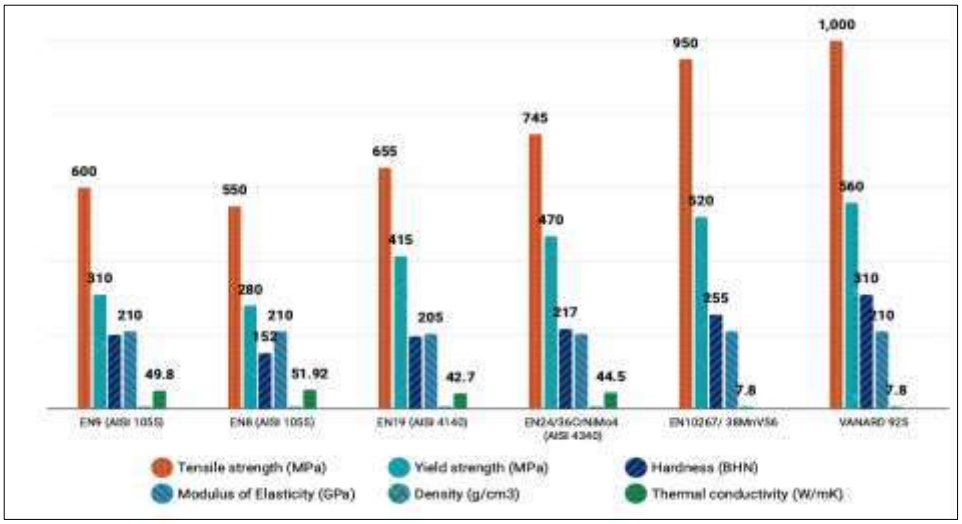


Figure 2: Comparative mechanical characteristics of particular crankshaft material.
Source: Authors, (2026).

II.2 PRODUCTION METHOD OF THE FORGED STEEL CRANKSHAFT

One of the most important steps in the production process is forging. It results in a product with higher mechanical qualities and less waste. In comparison to other methods of production, such as casting, this is a significant advantage [19]. Forging is a traditional method of producing high-performance or elevated crankshafts; it involves heating a billet to the proper forge temperature, and afterwards repeatedly hammering or pressing it between pairs of dies at quite high pressure. The dynamic qualities of a crankshaft, such as its resistance to impact, as well as cyclic loads, are improved during the forging process. The methodical procedure for producing a crankshaft [20] is shown in (Figure 3). Crankshafts that are forged are typically made by primarily using the processes of hot forging in combination with machining in order to put together the final product. [18-20] Monitoring the geometric diversity that results from shifts in temperature may be done in a number of different ways [21]. As the billet moves through the top and lower die configurations, it is purposefully distorted so that the ultimate product will be as good as it can possibly be [5], [22].

II.3 DEFECTS IN THE PRODUCTION OF CRANKSHAFTS DURING THE FORGING PROCESS

Operators and engineers may discover defects at any point after or during an activity or process and must speculate as to when, how, and why the defect emerges. At some point, researchers might require certain sources of evidence in order to predict the cause of the problem, the kind of part, the position of the defect on the component or part, the severity of the defect, etc. [23-24]. A framework of the defect diagnosis process is shown in (Figure 4) [25-26]. Forging defects are often the result of defective ingots, inaccurate heating, bad forging circumstances, inappropriate forging procedures, and unequal chilling of the stock upon forging [27]. In every forging operation, the manufacturer may run across a wide range of defects throughout the production process.

The product is often pulled off the market as a consequence of these problems. The defects in forged metallic components may be discovered with the use of image recognition, and the authors speculate that this method might be used to reduce stresses and fatigue throughout a physical assessment [28]. A cost-effective forged product free of forging defects is ensured by a well-designed shape and respect to the normal flow. Also among these factors are full cavity filling, good material properties, little die wear, longer die life, and good material properties [29]. To achieve an ideal design with a focus on stress distribution [1], forging defects must be eliminated, and metallic grain flow must be improved throughout the forging process.

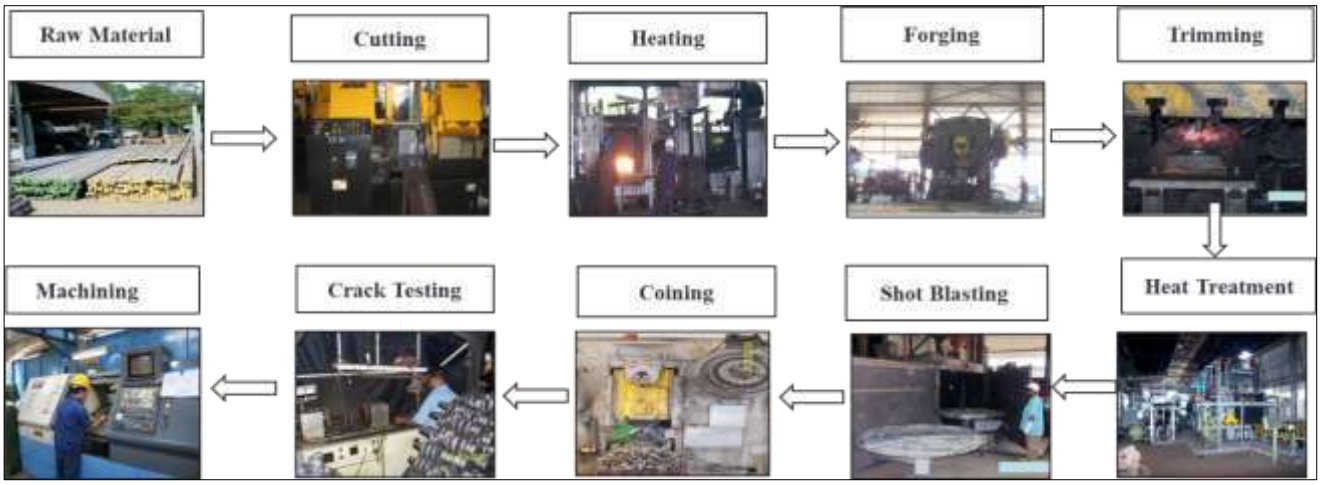


Figure 3: The Steps Involved in the Production of a Crankshaft.
Source: [19].

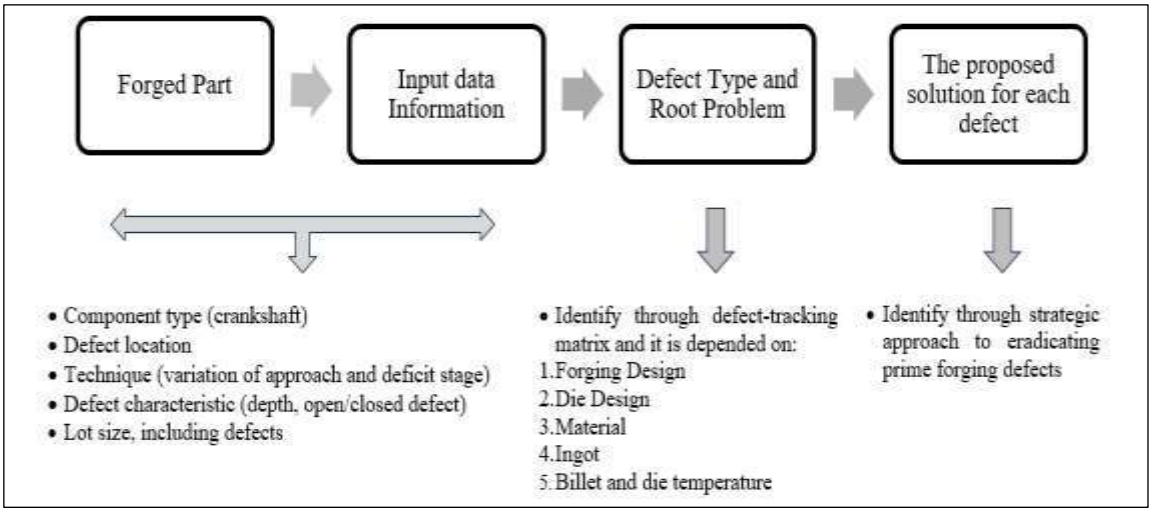


Figure 4: The basic forging defect. diagnosis process.
Source: [25-26]

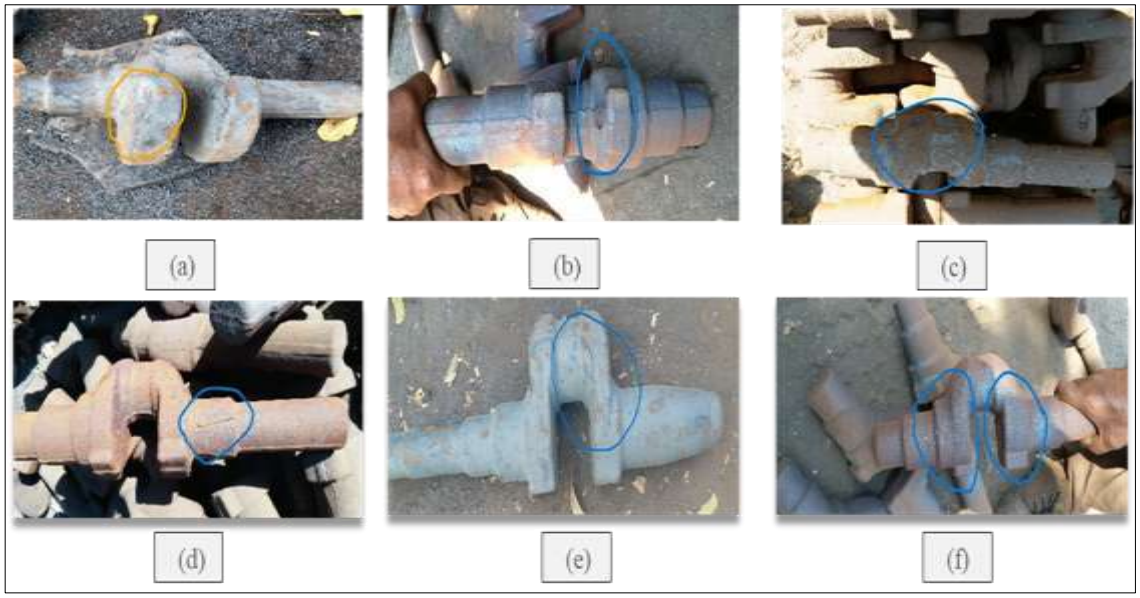


Figure 5: Forging Defects in Crankshafts(a) Unfills at counterbalance weight (b) lap at crankpin (c)Mismatch at side of journal (d) Mislocation on bearing area (e) Crack (f) Scale Pitting at on web and on bearing area.
Source: [19].

When it comes to forged components, the most typical cause of rejection or failure is the appearance of cracks on the inside [30]. The best strategy for decreasing forging defects was determined using the AHP-TOPSIS Method. The under-fill error is a significant issue in the field of investigation since it happens at such a high incidence in this sector [31]. Using the defect matrix, the first step entails the categorization of possible offenders into the following six groups: cracks, surface inadequacies, owing to shear flaws, folds, overlap defects, and form and structural defects. The second step classifies potential causes into different categories by using a tool called the defect matrix. This tool illustrates how different defects interact with each other as well as the factors that contribute to them [32]. The Deform-3d software application [33] can reliably forecast defects like underfill, residual stresses, and folds during fast heating. Defects are common in the crankshaft forging sector (Figures 5), with reflecting practices of recurrence [19].

II.4 METHODS AND STRATEGIES USED TO REDUCE REJECTION RATE IN CRANKSHAFT FORGING

This research article centers on crankshafts manufactured through hot forging, tailored for use in single-cylinder engines produced by Forge and Forge Pvt. Ltd. Rajkot. Industry faces an Approx. 5% rejection rate. Table 1 highlights production metrics over two consecutive 12-month periods: before corrective measures (2019-2020) and after (2020-2021) for the forging of single-cylinder crankshafts. Table 2 details defect types before and after implementing corrective measures. Table 3 outlines specific measures for preventing and addressing catastrophic major forging defects encountered in crankshaft production.

Table 1: Monthly Production Data and Defect Analysis Before and After Corrective Measures.

Metrics	Before Corrective Measures (12 Months)	After Corrective Measures (12 Months)
Total Production	120,284	113,112
Parts Accepted	114,783	110,927
Ultimate Rejection	5501	2185
Percentage of Final Rejection	4.57%	1.93%

Source: Authors, (2026).

Table 2: Defect Analysis and Contribution of each defect.

Defect Type	Counts, Before Corrective Measures (12 Months)	Counts, After Corrective Measures (12 Months)
Unfilling	3594	1565
Lap	716	188
Mislocation	521	229
Mismatch	355	122
Crack	73	44
Burnt	47	7
Scale Pit	177	30
Dent	11	0
Oversize	7	0
Total	5501	2185

Source: Authors, (2026).

Table 3: Preventing and correcting catastrophic forging defects in crankshafts.

Defect grading	Main forging defects	Protective and restorative measures
UF	Unfilling	Embed the blocker mode of operation of the dies by highlighting the areas that are not filled with the appropriate material.
		It is possible that the counterbalance masses for forged crankshaft may be revised or omitted altogether.
		Ensure that the employee receives the time and care they need in order to reduce the likelihood of mistakes occurring.
		Choose the kind of lubricant that is most suited for the specific die you are using and the corresponding raw material benchmark really has to be implemented as soon as possible.
		The integrity of the billet for the length of time provided may be easily confirmed by doing a physical check at many places more than thrice.
		Check the design of the roller and raise the depth by two to three millimetres in o rder to combat the underfilling effect.
		Set the burner towards the ideal temperature in order to heat it up enough.
LA	Overlap	To reduce the likelihood of incidents, provide a thorough training programme.
		In order to ensure consistency in quality, the right raw - materials norm must be implemented.
		Using a rubbing motion to grind the die's chamfered edges is advised.
		In order to fulfil the specifications of the design, the deformity of the blocker dies had to be eliminated.
		Before the user starts, the user should first preform the shape at the exact angle.

Source: Authors, (2026).

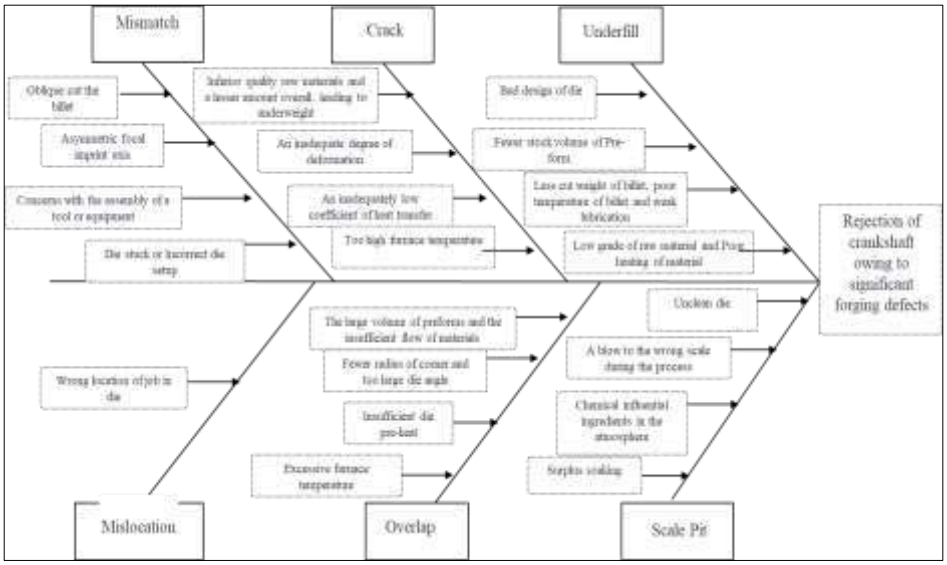


Figure 6: Cause effect of crankshaft rejection diagram for major defects of forging industry.
Source: Authors, (2026).

The cause-effect diagram (Figure 6) for significant defects in the forging sector, particularly with crankshaft rejections, depicts the main causes that contribute to incidences of rejection. The image above classifies significant defects and links their origins to certain root causes. By categorizing these components into distinct groups such as materials, processes, equipment, and human factors, it aids in the identification of possible causes of defects and provides direction for improvement initiatives. This tool is very beneficial for systematically assessing and correcting quality concerns in the forging process, with the goal of reducing rejection rates and improving overall product quality. Pareto analysis is a method used to discover and rank the most important factors that contribute to a challenge. Figure 7 illustrates the distribution of forging defects and percentages of rejection for single cylinder forged crankshafts in their original state, prior to the implementation of any remedial actions.

The defects are often classified according to their frequency or the extent to which may affect rejection rates. The highest bars in the Pareto chart represent the problems that occur most often or have the greatest severity. These bars suggest the areas where efforts should be concentrated in order to minimize rejection rates and enhance overall quality, prior to implementing remedial measures. The investigation reveals that over 80% of rejection difficulties stem from 20% of the detected faults. The Pareto principle implies that a vast majority of the rejections may be ascribed to a limited number of distinct fault categories. According to this research, the main reasons for rejecting single cylinder crankshafts forging are under-fills and overlaps, which account for over 75% of the rejections. This information is shown in Figure 7.

Figure 8 illustrates the outcomes of Pareto analysis after the implementation of remedial actions to resolve the highlighted defects from Figure 7. The image displays a revised proportion of defects and rates of rejection, demonstrating the influence of the implemented remedial measures. Optimally, after the implementation of measures such as process optimizations, quality control advancements, training initiatives, or equipment enhancements, the tallest bars in Figure 8 ought to indicate decreased frequencies or severities of faults. This indicates a positive advancement in reducing the number of rejected products and attaining superior product quality in the forging operation. According to the investigation, the main reasons for rejecting single cylinder crankshafts forging are under-fills and overlaps, which account for over 80% of the rejections. This information is shown in Figure 8.

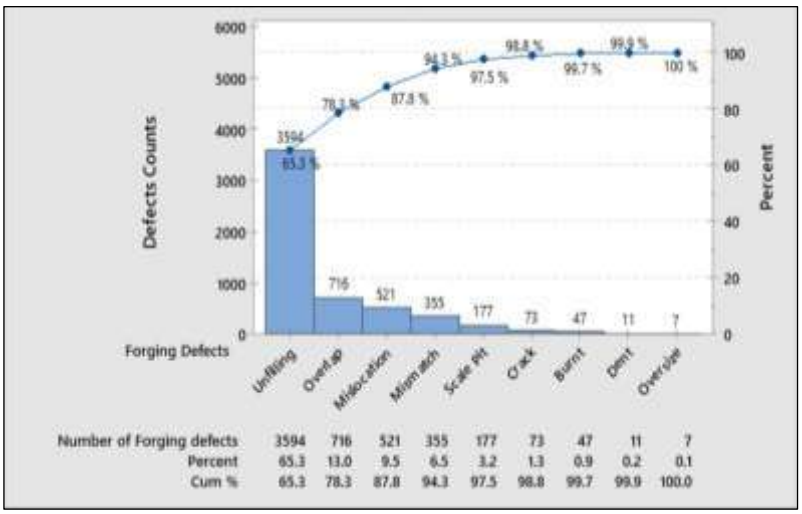


Figure 7: Pareto Diagram of Forging Defects counts and Rejection Rates for Forged Crankshafts
(Before Implementation of Corrective Measures)
Source: Authors, (2026).

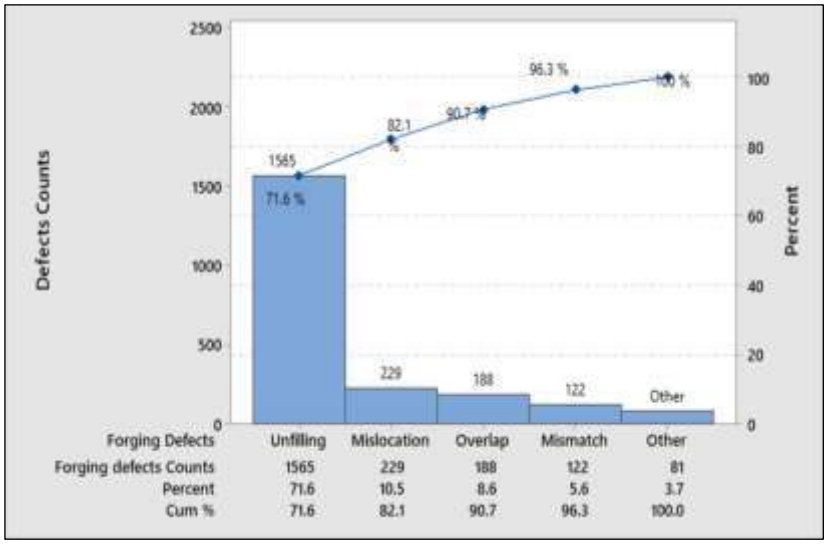


Figure 8: Pareto Diagram of Forging Defects counts and Rejection Rates for Forged Crankshafts (After Implementation of Corrective Measures).
Source: Authors, (2026).

III. RESULT AND DISCUSSIONS

III.1 DEFECT – TRACKING MATRIX FOR CRANKSHAFT FORGING INDUSTRY

An author [32] came up with the idea of a defect matrix, which is a tool that can be used to tie a forging defect first to factors that have the potential to cause it. A cause-effect diagram is a visualization tool that may be used to logically arrange the potential causes of a certain issue or repercussion by visually portraying those potential causes in greater detail. It can also reveal causal links between different hypotheses. A fishbone diagram, or an Ishikawa diagram, is one of the most common names for this form. [34-36] Figure 11 is a cause-and-effect diagram that illustrates the massive defects that led to the crankshaft rejection during the manufacturing process. The design and arrangement of a defect-tracking matrix is shown in (Figure 9-11), which is relevant to the crankshaft forging industrial sector.

The Defect-Tracking Matrix has twenty-two flaws arranged vertically and upwards of five groups (Table 4) of probable causes for the defects arranged horizontally. The major groups are presented just to make the matrix simple to read; nevertheless, in (Figure 9-11), a portion of the whole Defect-Tracking Matrix or Cause Effect Matrix is shown; this section contains the characterization of the self-styled foundation as like specialist assistance for forging. In (Figure 9-11) the symbol "*" shows that there is a link between both the main causes or possible factors contributing to rejections and forging defects. (Figure 9-11) depicts the second defect tracking matrix for the crankshaft forging sector in accordance with the operator and materials factors. (Figure 9-11) depicts the first defect tracking matrix, or cause and effect matrix, for the crankshaft forging enterprise in terms of tool and machine factors.

(Figure 9-11) depicts the second defect tracking matrix for the crankshaft forging sector in accordance with the methods and environmental factors. The marks in the defect-tracking-matrix are dependent on the process that is currently being used, which is the reason why a cause-effect matrix is required to be developed for each and every type of forging process. A task of this kind will need much too much time and will be very challenging to summarise. Because of this, the processes need to be separated into their respective primary groups according to criteria such as the geometry of the tool and workpiece or the operation that is more frequent. For each Defect-Tracking-Matrix, the marks that indicate the link between the defect and what might cause it will produce an agenda that is unique to the process that is now being carried out.

Table 4: Five groups' Parameters for possible causes laid out horizontally.

Vertically Forging Defects	**Horizontally Probable causes**				
	Main Group 1 Process	**Main Group 2 Environmental**	**Main Group 3 Tools**	**Main Group 4 Materials**	**Main Group 5 Operators**
Example Laps	Sub group 1 Sub group 2	Sub group 1 Sub group 2	Sub group 1 Sub group 2	Sub group 1 Sub group 2	Sub group 1 Sub group 2
	Each Sub group Cause 1 Cause 2	Each Sub group Cause 1 Cause 2	Each Sub group Cause 1 Cause 2	Each Sub group Cause 1 Cause 2	Each Sub group Cause 1 Cause 2

Source: Authors, (2026).

III.2 EFFECTIVENESS OF CORRECTIVE MEASURES IN REDUCING DEFECTS AND IMPROVING QUALITY

Based on the data analysis presented in Tables 1, 2, and insights from Table 3 regarding corrective measures, significant improvements in production quality and defect reduction are evident following the implementation of strategies at Forge and Forge Pvt. Ltd. Rajkot. Over a 12-month period, total production slightly decreased from 120,284 to 113,112 parts and the acceptance rate improved from 114,783 to 110,927 parts. Notably, ultimate rejections dropped from 5,501 to 2,185 parts, with final rejection rates decreasing from 4.57% to 1.93%. Specific defect reductions include unfills (3,594 to 1,565 occurrences) and lap defects (716 to 188 instances). These improvements reflect effective strategies like die design optimization, process controls, and operator training, enhancing quality, customer satisfaction, and operational efficiency in the competitive forging industry.

Defect Tracking Matrix for Crankshaft Forging Industry sectors (Cause Effect Matrix)			Possible Factors Contributing to Rejections																		
			Parameters for Tool/Workpiece Material										Parameters for Operators								
			Stress Crack/Ingot Material Defect) of Billet	Too Low Heat Transfer Coefficient	Too High Thermal Expansion Coefficient	Inclusions	Underweight	Discontinuities in Microstructure	Too Soft Die Material	Too High Elastic Material	Poor Raw Material Grade	Too Large Grains in the Material	Lack of Training	Less Skilled Worker	Improper Placement of Stock on Die	Improper Die setting	Surplus Material escapes to side	Improper Scale Blowing in Operation	Careless Handling of Jobs	Uncleaned Dies (To Soil with dirt or impurities)
	Sr.No.	Forging Defect	Code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Forging Defects	1	Mismatch	MM								*			*	*		*		*		
	2	Unfilling	UF	*				*		*		*		*	*		*			*	
	3	Closure	CL																		
	4	Oversize	OS							*											
	5	Undersize	US			*															
	6	Mislocation	ML																		
	7	Overlap	LA	*										*	*		*		*	*	
	8	Chopping	CH														*				
	9	Overheating	OH																		
	10	Dent Mark	DM													*				*	
	11	Punch Mark	PM																		
	12	Bend	WA																		
	13	Twist	TW																		
	14	Surface Crack	SC	*	*	*		*	*			*								*	
	15	Eccentricity	EC								*										
	16	Scale Pit	SP				*					*							*		*
	17	Excess Burr	EB							*		*			*			*			
	18	Excess Shear	ES																		
	19	Poor Mechanical Properties	MP								*										
	20	Poor Formability	FL				*														
	21	Monogram	UM																		
	22	Heat Number	UH																		

Figure 9: Defect Tracking Matrix of Tool Materials and Employee Constraints for the Crankshaft Forging Sector.
Source: Authors, (2026).

Defect Tracking Matrix for Crankshaft Forging Industry sectors (Cause Effect Matrix)			Possible Factors Contributing to Rejections																						
			Parameters for Machines/Tools																						
			Preform						Design & Geometry												Press		Surface		
			Less Stock Volume of Preform	Too Large Volume of Preform	Improper Preform Geometry	Eccentric Centre Line of Impression	Obliquely Cut Billet	Narrow Instable Sections	Poor Die Design	Insufficient Die Preheat	Misplacement between Dies	Blocker Mismatch	Inappropriate and Poor Geometry of Die	Die Locked	Wrong Location of part in Tool	Less corner Radius	Too Large Die Angle	Sharp Edges of Tool	Tool Assemblings Error	Incorrect Positioning of Flash	Inadequate Press Force Energy and/or Power	Improper Capacity	Die to Die Land Worn Out	Punch Worn Out	Surface Worsen of Die	Tool Hardness	Too Rough Surface
	Sr.No.	Forging Defect	Code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Forging Defects	1	Mismatch	MM				*	*					*		*					*			*				
	2	Unfilling	UF	*		*				*	*			*								*						*
	3	Closure	CL							*														*				
	4	Oversize	OS																							*		
	5	Undersize	US																									
	6	Mislocation	ML													*												
	7	Overlap	LA		*	*		*	*	*	*			*			*	*	*									
	8	Chopping	CH									*																
	9	Overheating	OH																									
	10	Dent Mark	DM																									
	11	Punch Mark	PM																						*			
	12	Bend	WA																									
	13	Twist	TW							*																		
	14	Surface Crack	SC					*										*										
	15	Eccentricity	EC					*												*								
	16	Scale Pit	SP																									
	17	Excess Burr	EB													*											*	
	18	Excess Shear	ES		*	*											*				*							
	19	Poor Mechanical Properties	MP																									
	20	Poor Formability	FL																									
	21	Monogram	UM																									
	22	Heat Number	UH																									

Figure 10: Defect Tracking Matrix of Machine/Tool Constraints for the Crankshaft Forging Sector.
Source: Authors, (2026).

Defect Tracking Matrix for Crankshaft Forging Industry sectors (Cause Effect Matrix)

				Possible Factors Contributing to Rejections																					
				Parameters of Process																				Parameters for Environmental		
				Thermic				Due To Flow				Geometric					Local			Friction					
				The temperature of the furnace is too high.	Less Temperature of Billet	Heating at a very rapid rate	Poor Heating of Material	Improper Material Flow	Insufficient Number of Blows	Low Cut Weight of Billet	Improper Deformation	Heterogeneous Deformation	Incorrect Position of Billet	Too protracted a stroke	Distortion along the Length of Part	Distortion Across the Width of Part	Improper Arrangement After Cutting	Monogram not Visible	Heat Lot not Visible	Too High Friction	Low grade Lubricant	Improper Lubrication	The Presence of Chemically Active Contaminants in the Environment	Excess Soaking	Surplus Air in the Surrounding Environment
	Sr.No.	Forging Defect	Code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Forging Defects	1	Mismatch	MM										*												
	2	Unfilling	UF	*	*		*	*	*	*		*	*									*	*			
	3	Closure	CL	*																					
	4	Oversize	OS																						
	5	Undersize	US																						
	6	Malocation	ML																						
	7	Overlap	LA	*	*			*					*	*						*	*	*				
	8	Chopping	CH																						
	9	Overheating	OH	*																		*			
	10	Dent Mark	DM																						
	11	Punch Mark	PM																						
	12	Bend	WA												*		*								
	13	Twist	TW													*									
	14	Surface Crack	SC	*		*	*				*			*										*	
	15	Eccentricity	EC										*												
	16	Scale Pit	SP	*																		*	*	*	*
	17	Excess Burr	EB																						
	18	Excess Shear	ES											*						*					
	19	Poor Mechanical Properties	MP																						
	20	Poor Formability	FL																						*
	21	Monogram	UM															*							
	22	Heat Number	UH																*						

Figure 11: Defect Tracking Matrix of Process and Environmental Constraints for the Crankshaft Forging Sector.
Source: Authors, (2026).

IV. CONCLUSIONS

1. The Defect-Tracking-Matrix or Cause-Effect Matrix may provide guidance to the manufacturing and design specialists, allowing them to resolve a problem with the product and create a robust component. The foundation of a Defect-Tracking-Matrix was created for the economic advantage of the crankshaft forging industrial sectors to correct forging defects. The developer may get direction from the cause- effect matrix using any of two approaches. The first approach is referred to as the "detection mode," and it's used when the developer wants to figure out why a certain defect is present in the product and exactly how to get rid of it. Because the Defect-Tacking Matrix is being used horizontally throughout this context, it is feasible, for a given defect, to acquire data on which variables might produce the defect as well as a potential strategy to avoid it. This is because the defect-tracking-matrix is used in this context.

The 'advising solutions mode' of the Cause Effect Matrix is another use of its capabilities. With the 'advising solutions mode' method, the Defect-Tracking-Matrix is used in a vertical orientation, providing information on the defect possibility associated with each operating factor. It presents the theoretical foundation for a method to reduce the overall quantity of defects in a forged product. The Defect-Tracking - Matrix is a database or catalogue of defects that may be generated for a wide range of forging methods and workpiece shapes to serve as a helpful tool for the design process. It is possible for a design engineer to employ a Defect-Tracking-Matrix in practice, although the factors of that procedure may not match those given here. Because of this, more tests using numerical or parametric methods may be needed to back up the Defect-Tracking-Matrix results.

2. The implementation of targeted corrective measures at Forge and Forge Pvt. Ltd. Rajkot effectively reduced total rejection rates and specific major forging defects, significantly enhancing production quality and operational efficiency.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Dr. Amitkumar B. Solanki
Methodology: Dr. Amitkumar B. Solanki
Investigation: Dr. Amitkumar B. Solanki
Discussion of results: Dr. Amitkumar B. Solanki
Writing – Original Draft: Dr. Amitkumar B. Solanki
Writing – Review and Editing: Dr. Amitkumar B. Solanki, Dr. Nisha Bora, Dr. Nilesh H. Pancholi
Resources: Dr. Amitkumar B. Solanki, Dr. Nisha Bora, Dr. Nilesh H. Pancholi and Prof. Nilesh R. Tank
Supervision Dr. Amitkumar B. Solanki, Dr. Nisha Bora, Dr. Nilesh H. Pancholi
Approval of the final text: Dr. Amitkumar B. Solanki, Dr. Nisha Bora, Dr. Nilesh H. Pancholi and Prof. Nilesh R. Tank

VI. REFERENCES

[1] M. D. V. Kumar and U. V. Kumar, “Defect rectification in forging operation and optimizing the process parameters,” in Lecture notes in mechanical engineering, 2018, pp. 533–542. doi: 10.1007/978-981-13-1724-8_49.

[2] D. Kumar and S. Mondal, “A review on modelling,design and optimization of forging process,” IOP Conference Series Materials Science and Engineering, vol. 1126, no. 1, p. 012001, Mar. 2021, doi: 10.1088/1757-899x/1126/1/012001.

[3] K. Lange, Handbook of Metal Forming, New ed. Dearborn, MI, USA: Society of Manufacturing Engineers, 1994.

[4] Ü. Şengül and A. B. Şengül, “Hot forging die material selection using fuzzy multi-criteria decision-making methods,” Materials Today Communications, vol. 38, p. 108352, Feb. 2024, doi: 10.1016/j.mtcomm.2024.108352.

[5] American Society for Metals, Metals Handbook: Forming and Forging, 9th ed. Materials Park, OH, USA: ASM International, 2017.

[6] T. G. Byrer, S. L. Semiatin, & D. C. Vollmer. Forging Handbook (2nd edition), American Society for Metals, 1985.

[7] S. Asiri, “Modeling and Analysis of Automotive Engine Crankshaft Made of Composite and Functionally Graded Materials.,” Advances in Materials Science and Engineering, vol. 2022, pp. 1–15, Apr. 2022, doi: 10.1155/2022/4005368.

[8] M. Degefe, P. Paramasivam, T. Dabasa, and S. V. Kumar, “Optimization and finite element analysis of single cylinder engine crankshaft for improving fatigue life,” American Journal of Mechanical and Materials Engineering, vol. 1, no. 3, p. 58, Jun. 2017, doi: 10.11648/j.ajmme.20170103.11.

[9] F. Hussain, “Dynamic Structural Analysis of Engine Crankshaft at Different Angle of Crank Turns for Three Different Materials.,” Journal of Mechanics of Continua and Mathematical Sciences, vol. 16, no. 3, Mar. 2021, doi: 10.26782/jmcms.2021.03.00007.

[10] L. Karthick, N. Mallireddy, J. Yogaraja, S. Sivakumar, “Modelling and Analysis of an EN8 crankshaft material in comparison with Forged steel crankshaft,” Materials Today Proceedings, vol. 47, pp. 6168–6172, Jan. 2021, doi: 10.1016/j.matpr.2021.05.079.

[11] Y. Wang et al., “Failure analysis and improvement of a 42CrMo crankshaft for a heavy-duty truck,” Engineering Failure Analysis, vol. 153, p. 107567, Aug. 2023, doi: 10.1016/j.engfailanal.2023.107567.

[12] L. R. Kumar, A. Ravinthiran, P. Srivarshani, S. Sharadha, V. Saravanan, and R. A. Kumar, “Investigations on multi cylinder four stroke diesel engine’s crankshafts with different materials,” Materials Today Proceedings, Mar. 2023, doi: 10.1016/j.matpr.2023.03.058.

[13] A. D. Pearson, Ed., SAE Ferrous Materials Standards Manual. Warrendale, PA, USA: Society of Automotive Engineers, 2004.

[14] P. D. Harvey, Engineering Properties of Steels. Metals Park, OH, USA: American Society for Metals, 1982.

[15] H. S. Nguyen et al., “Selection of crankshaft manufacturing material by the PIV method,” Engineering Technology & Applied Science Research, vol. 14, no. 4, pp. 14848–14853, Aug. 2024, doi: 10.48084/etasr.7514.

[16] J. R. Davis, ASM Specialty Handbook - Carbon and Alloy Steels. Materials Park, OH, USA: Davis & Associates, ASM International, 1996.

[17] J. G. La Barbera-Sosa et al., “Impact of local forging parameters on the microstructural evolution and mechanical properties of hot forging steel dies,” Materials Science and Technology, May 2025, doi: 10.1177/02670836251346513

[18] S. K. Jha, "Cost optimization of a forged steel crankshaft in a global manufacturing environment," Int. J. Mech. Eng. Res. Technol., vol. 2, no. 2, pp. 16–28, 2016

[19] Forge & Forge Pvt. Ltd. (n.d.). Home Page. Retrieved from https://www.forgenforge.com/forging.html

[20] P. Ruban and J. Jayaprakash, “Implementation of design for manufacturing in hot forging,” IOP Conference Series Materials Science and Engineering, vol. 992, no. 1, p. 012011, Nov. 2020, doi: 10.1088/1757-899x/992/1/012011.

[21] V. Bhoyar and S. Umredkar, "Manufacturing processes part II: A brief review on forging," Int. J. Innov. Eng. Sci., vol. 5, pp. 25–32, Jan. 2020.

[22] L. F. Moreira et al., “O uso das ferramentas da qualidade para a resolução de não conformidades em produtos no setor moveleiro do Rio Grande do Sul,” Revista De Gestão E Secretariado (Management and Administrative Professional Review), vol. 15, no. 1, pp. 931–954, Jan. 2024, doi: 10.7769/gesec.v15i1.3398.

[23] Y. Ma et al., “Using deep reinforcement learning for zero defect smart forging,” in Advances in transdisciplinary engineering, 2022. doi: 10.3233/atde220189.

[24] E. S. Gharat, "Application of Industry 4.0 in forging operations," J. Emerg. Technol. Innov. Res., vol. 7, no. 1, pp. 26–36, 2021.

[25] S. Mewada et al., “Smart Diagnostic Expert System for defect in forging process by using machine learning process,” Journal of Nanomaterials, vol. 2022, no. 1, Jan. 2022, doi: 10.1155/2022/2567194.

[26] D. Guo et al., “Synchroperation in industry 4.0 manufacturing,” International Journal of Production Economics, vol. 238, p. 108171, May 2021, doi: 10.1016/j.ijpe.2021.108171.

[27] Y. Liu, Y. Wu, J. Wang, and S. Liu, “Defect analysis and design optimization on the hot forging of automotive balance shaft based on 3D and 2D simulations,” The International Journal of Advanced Manufacturing Technology, vol. 94, no. 5–8, pp. 2739–2749, Sep. 2017, doi: 10.1007/s00170-017-1080-9.

[28] T. Yamazaki and A. Fukui, “Defect detection for forged metal parts by image processing,” International Journal of Future Computer and Communication, vol. 9, no. 1, pp. 23–26, Mar. 2020, doi: 10.18178/ijfcc.2020.9.1.560

[29] T. Quazi and M. M. Naikwade, "Review of forging defect and its remedies," Trends Mech. Eng. Technol., vol. 10, no. 2, pp. 1–3, 2020.

[30] Jiang, B, B., Dong, Z. q., Yang, Z., Zhou, L. y., Liu, Y. z., & Wang, Y. n. 2015. Analysis of the Formation of Surface Crack on Crankshaft After Die Forging. Transactions of the Indian Institute of Metals, 68(4), 553–559. https://doi.org/10.1007/s12666-014-0485-5

[31] A. Panwar, R. Kumar, R. Thakur, B. Goel, A. Rana, A. Pathania, and V. Aggarwal, "Selection of optimal parameters for reduction of forging defect using AHP-TOPSIS technique," Int. J. Emerg. Technol., vol. 11, no. 2, pp. 178–186, 2020.

[32] M. Arentoft and T. Wanheim, “The basis for a design support system to prevent defects in forging,” Journal of Materials Processing Technology, vol. 69, no. 1–3, pp. 227–232, Sep. 1997, doi: 10.1016/s0924-0136(97)00023-x.

[33] N. Jose, S. S. Kumar, S. Ku, J. Varghese, and S. Joshy, "Stress Analysis and Reducing Rejection Rate of Work in Forging Using Deform-3d.," Int. J. Eng. Adv. Technol., vol. 10, no. 4, pp. 1–7, 2020. doi: 10.9790/9622-1004040107.

[34] T. P. Bagchi, Total quality management: tools and implements. IE(I) Journal, 78,1997.

[35] S.E. Burke, & T.R. Silvestrini, The Certified Quality Engineering Handbook, 4th ed, ASQ Quality Press, 2017.

[36] D. Summers, Quality, 6th ed. Columbia, SC, USA, 2017.

Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 1229-1239. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3152

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

ARTIFICIAL NEURAL NETWORK BASED PREDICTIVE PERFORMANCE ANALYSIS OF PHOTOVOLTAIC OUTPUT UNDER VARIOUS SOLAR RADIATION RANGES

Ravshanbek Rakhmatulaev*[1], Rodney Tan[2], Aliev Rayimjon[3], Yu L. J.[4]

[1, 4]Department of Mechanical Engineering, Faculty of Engineering, UCSI University, 56000 Kuala Lumpur, Malaysia.
[2]Department of Electrical and Electronic Engineering, Faculty of Engineering, UCSI University, 56000 Kuala Lumpur, Malaysia.
[3]Department of Physics, Faculty of Physics and Mathematics, Andijan State University, Andijan 170100, Uzbekistan.

[1]http://orcid.org/0009-0008-4799-5689, [2]http://orcid.org/0000-0002-2364-6088
[3]http://orcid.org/0000-0002-7375-727X, [4]http://orcid.org/0000-0002-8195-1566

Email: *rraxmatullayev11@gmail.com, rodneytan@ucsiuniversity.edu.my, alievuz@yahoo.com, yulj@ucsiuniversity.edu.my

ARTICLE INFO

Article History
Received: December 16, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Solar energy
Photovoltaic power
Prediction
FFBPNN

ABSTRACT

Accurate assessment of photovoltaic (PV) power output under varying environmental conditions is essential for evaluating solar energy performance and optimizing its application. This study employs a Feedforward back propagation neural network (FFBPNN) approach to predict the output power of a solar panel using a combination of electrical and weather-related parameters. The input variables include short-circuit current (I_{sc}), open-circuit voltage (V_{oc}), maximum voltage (V_{max}), maximum current (I_{max}), efficiency (EFF) - represents how effectively it converts incident solar energy into electrical energy., fill factor (FF) - the ratio of the maximum obtainable power to the product of the open-circuit voltage and short-circuit current, solar radiation (G), wind speed (v), ambient temperature (T_{am}), and panel temperature (T_{panel}) under solar radiation intensity in the ranges of 200–400, 400–600, 600–800, and 800–1000 W/m². Three artificial neural network (ANN) models, ANN-10 (10 inputs), ANN-8 (8 inputs, except I_{sc} and V_{oc}), and ANN-6 (6 inputs, except I_{sc}, V_{oc}, V_{max} and I_{max}), are developed for 496 dataset and compared based on their prediction accuracy using the coefficient of correlation (R) by varying a number of neurons in the hidden layer from 1 to 15. Among them, the ANN-8 model (8-10-1) demonstrates the highest prediction accuracy (R value of training, testing, validation and overall model exceeding 0.999) particularly at high solar radiation levels (800–1000 W/m²). The results confirm that FFBPNN-based modelling is a reliable and effective tool for forecasting PV output power under diverse operating conditions. This approach supports solar energy potential assessment and performance optimization across various geographical and climatic contexts.

Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

In response to growing global energy demand, environmental degradation, and the urgent need to reduce reliance on fossil fuels, solar energy has become a cornerstone of sustainable development. Among renewable energy sources, solar power is particularly attractive due to its abundance, scalability, and minimal environmental impact [1]. Photovoltaic (PV) technology, which converts sunlight directly into electricity, plays a central role in harnessing this resource [2]. The increasing integration of PV systems across residential, commercial, and utility-scale applications reflects a global shift toward clean energy. However, the intermittent and variable nature of solar irradiance, shaped by geographic, meteorological, and temporal factors, poses significant challenges to the consistent and efficient operation of PV systems [3]. Accurate forecasting and modelling of PV output are thus critical for grid stability, energy planning, and operational efficiency [4].

PV system performance is governed by complex, nonlinear relationships between environmental parameters (e.g., solar irradiance, temperature, wind speed, humidity) and electrical outputs (e.g., voltage, current, power) [5]. Predicting output under varying conditions is vital not only for system monitoring and control but also for optimizing energy storage, demand response, and economic dispatch, especially in smart grids and hybrid energy systems. Despite its importance, accurately modelling PV behaviour remains difficult due to the stochastic and nonlinear nature of these influencing factors [6]. Traditional modelling approaches, empirical, analytical, and physics-based, have laid the foundation for PV system analysis. These methods often rely on simplified assumptions and require parameter calibration, which may not generalize well under dynamic real-world conditions [7]. For instance, the widely used single-diode model requires parameters that could be hard to measure or that vary over time. Similarly, regression-based methods assume fixed functional relationships, limiting their ability to capture the nonlinear dynamics of PV performance. These limitations highlight the need for more adaptable, data-driven approaches [8].

In this context, machine learning (ML) has emerged as a promising alternative for modelling complex systems, including PV output prediction. Among ML techniques, Artificial Neural Networks (ANNs) are particularly well-suited for this task due to their ability to model nonlinear relationships, manage noisy data, and generalize from historical patterns [9]. Modelled after the human brain, ANNs comprise layers of interconnected neurons that transform input features through weighted connections and activation functions to generate outputs. This structure allows ANNs to approximate complex functions without requiring predefined mathematical models [10]. The application of ANNs to PV modelling represents a shift from equation-driven to data-driven methodologies [11]. Leveraging large volumes of operational and meteorological data, often collected by sensors and monitoring systems, ANNs adapt to different PV technologies, configurations, and climate conditions. They could also be integrated with other intelligent systems such as fuzzy logic, genetic algorithms, or hybrid ML frameworks to further improve prediction accuracy and robustness [12]. Numerous studies have validated the effectiveness of ANNs in predicting PV performance [13- 15].

For instance, a Levenberg-Marquardt back-propagation ANN model with six input parameters and a hidden layer of up to 35 neurons achieved high accuracy in output prediction [16]. Other studies compared radial basis and feed-forward topologies using different training algorithms, finding mean bias errors below 1% and prediction errors between 0.05% and 1% [17-20]. In another case, an MLP model with seven inputs predicted maximum voltage (V_m) and current (I_m) with high precision using a single hidden layer of up to ten neurons [21]. A feed-forward ANN with two hidden layers, trained on six meteorological variables, achieved R^2 values as high as 99.9% when predicting mono-crystalline PV output [22]. Mellit et al. (2013) demonstrated that even simple ANN structures could effectively predict PV output. Using one year's worth of data for a 50 W_p module, their model, with just a single hidden neuron, achieved determination coefficients of 96–97% on sunny days and 93–97% on cloudy days [23]. Compared to traditional techniques, ANN-based models consistently offer superior accuracy in capturing the effects of environmental variability, shading, and system degradation.

They also provide practical benefits in terms of implementation speed and computational efficiency, especially with modern hardware and software capabilities [24]. Nevertheless, the effectiveness of ANN models depends on factors such as network architecture, input feature selection, training algorithms, and data quality [25]. Despite their promise, ANN-based models in PV energy forecasting often face challenges such as overfitting, convergence instability, and limited interpretability. Recent advances, however, are steadily overcoming these limitations. Regularization strategies, ensemble learning, and data augmentation enhance generalization, while adaptive optimizers, metaheuristic-assisted training, and advanced initialization improve convergence speed and stability. Furthermore, explainable AI tools, attention mechanisms, and hybrid physics-informed networks enhance transparency and trust in predictions. Together, these developments expand the reliability and applicability of ANN models, enabling more accurate forecasting, fault detection, and system optimization, thereby strengthening their role in modern PV energy research and grid integration [26]. Despite the demonstrated success of ANN-based models in PV output prediction, several research gaps remain.

Many existing studies focus on limited datasets, specific PV technologies, or narrow environmental conditions, which restrict the generalizability and robustness of the models [24], [26]. Furthermore, there is a lack of systematic comparison between different ANN architectures in terms of prediction accuracy, training efficiency, and sensitivity to input features. Additionally, few studies explore the integration of real-time meteorological data or assess model performance across diverse climatic zones [25]. These limitations highlight the need for more comprehensive, adaptable, and scalable ANN frameworks tailored to dynamic, real-world conditions. The present study aims to address these gaps by developing a data-driven ANN model that leverages diverse environmental inputs and evaluates multiple neural network configurations to optimize PV power output prediction. The novelty of this work lies in its systematic exploration of ANN architectures under varying climatic conditions, the inclusion of a broader set of input features, and the evaluation of model performance across different PV technologies. This approach not only enhances prediction accuracy but also contributes to the development of more reliable and transferable forecasting tools for modern energy systems.

Hence, this study aims to develop and evaluate an ANN model for accurately predicting the output power of photovoltaic (PV) systems under varying environmental conditions. The primary objective is to investigate the impact of different electrical and meteorological input parameters on the accuracy of PV power prediction. Specifically, the study seeks to: (i) assess the predictive performance of ANN models using a combination of electrical parameters (short-circuit current (I_{sc}), open-circuit voltage (V_{oc}), maximum voltage (V_{max}), maximum current (I_{max}), efficiency (EFF), fill factor (FF)) and environmental factors (solar radiation, wind speed, ambient temperature, and panel temperature); (ii) analyse the effect of different solar radiation intensity ranges on model accuracy; (iii) optimize ANN architecture by varying the number of input features and hidden layer neurons; and (iv) compare the prediction accuracy of three ANN models (ANN-10, ANN-8, and ANN-6) using the coefficient of correlation (R). The findings aim to support the application of ANN-based modelling for PV performance evaluation and solar energy forecasting across diverse climatic conditions.

II. METHODOLOGY

This study develops an ANN model to predict the output power of a photovoltaic (PV) module under varying environmental conditions. The model is trained and evaluated using field experimental data collected over a one-month period at UCSI University Field Campus, Kuala Lumpur, Malaysia (Latitude: 3.079548° N, Longitude: 101.733216° E, Elevation: 70 m). The site experiences a tropical climate, making it suitable for solar energy potential assessment and performance prediction.

II.1 PV MODULE SPECIFICATIONS AND MEASUREMENT INSTRUMENTS

The solar panel model ZDNY-250C60 is made of Monocrystal (a-Si) material. The panel size is 1650×992×40 mm, and it consists of 60 cells. Its maximum power output is 250 W. The open-circuit voltage (V_{oc}) is 37.85 V, the short-circuit current (I_{sc}) is 8.40 A, the voltage at maximum power (V_{max}) is 31.17 V, and the current at maximum power (I_{max}) is 8.03 A. The maximum system voltage is 1000 V, and the maximum fuse current is 15 A. The output tolerance is 0/+3%, and the measurements were taken at a temperature of 25°C. Each cell has a dimension of 125×125 mm Figure (1).

II.2 DATA COLLECTION AND INPUT FEATURES

Data were collected using the experimental setup shown in Fig.(1), over a month period from 25 June to 26 July 2025, with measurements taken daily between 9:00 a.m. and 5:00 p.m. A PROVA 200A I-V Tracer Solar Module Analyzer was used to measure Open circuit voltage (V_{oc}), Short-Circuit current (I_{sc}), Voltage at max power (V_{mp}), and Current at max power (I_{mp}) capture the electrical characteristics of the PV module. The surface temperature of the panel was measured using a FLUKE 59 Mini IR Thermometer, while ambient temperature was recorded via a wired Mercury thermometer. Wind speed was monitored using a handheld anemometer. The following parameters were recorded at regular intervals: (i) Electrical parameters: I_{sc}, V_{oc}, V_{max}, I_{max}, Efficiency (EFF), Fill Factor (FF), (ii) Environmental parameters: Solar Radiation (G), Wind Speed (v), Ambient Temperature (T_{am}), Panel Temperature (T_{panel}), and (iii) The PV output power served as the target (output) variable for the ANN model. To evaluate model performance under different irradiance conditions, solar radiation data were grouped into four categories: 200–400, 400–600, 600–800, and 800–1000 W/m².

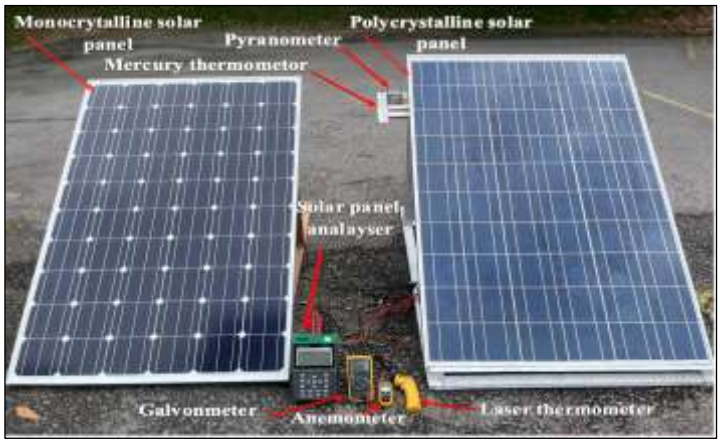


Figure 1: Experimental set-up used in the present study.
Source: Authors, (2026).

II.3 ANN MODELLING

Artificial Neural Network (ANN) is a computational modelling framework inspired by the structure of the human brain, consisting of interconnected processing units (neurons) arranged in three layers: input, hidden, and output [27]. In this study, a feed-forward Multilayer Perceptron (MLP) architecture was employed Figure (2), implemented in MATLAB 2024. Each input parameter is supplied to a neuron in the input layer, processed through synaptic weights and activation functions within the hidden layer, and finally transformed into the predicted output in the output layer [28].

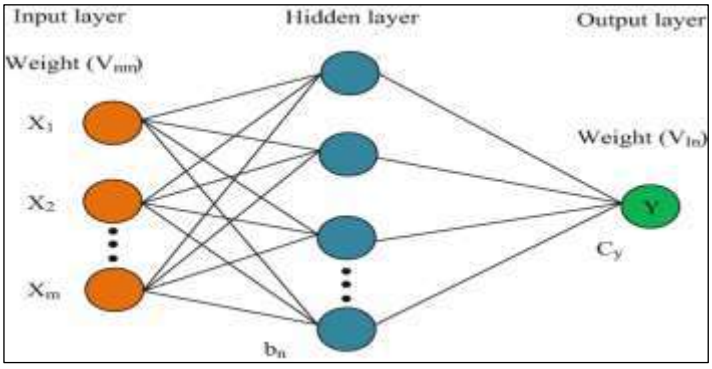


Figure 2: General architecture of ANN.
Source: Authors, (2026).

The dataset containing 40 entries was divided into 70% training, 15% testing, and 15% validation. Before training, all input parameters were normalized to a [0, 1] range using the Eq. (1):

$$X_i = \frac{0.8}{d_{max} - d_{min}} \left(d_j - d_{min}\right) + 0.1 \quad (1)$$

Where d_{max}, d_{min}, and d_i are the maximum, minimum, and current values of the dataset, respectively. The ANN model was trained using the Levenberg–Marquardt algorithm, known for its fast convergence. The common activation functions used in ANN are provided in Eq. (2). The sigmoid activation function was used in the hidden layer to introduce nonlinearity and enable the ANN to learn the complex nonlinear relationships inherent in solar radiation and PV electrical behaviour. The output layer used a linear activation function because the target variable (PV output power) is continuous and unbounded, requiring an unrestricted output range. This sigmoid–linear combination is the standard architecture for regression tasks and ensures both stable training and accurate numerical prediction.

$$f(s) = \begin{cases} S\ linear\ function \\ \frac{1}{1+e^{-s}}\ \ sigmoid\ function \\ \frac{e^{-s}}{e^{+s}+e^{-s}}\ \ hyerbolic\ tangent\ function \end{cases} \quad (2)$$

The error function, as given in Eq. (3), to be minimized during training is given in by:

$$E = \frac{1}{2}\sum_{k=1}^{l}(\overline{y_k} - y_k)^2 \quad (3)$$

The ANN models were evaluated using Coefficient of Correlation (R), which measures the strength of the linear relationship between predicted and actual output. To analyse the influence of different input combinations(in shown Table 1), three ANN models were developed:

Table 1: Different input combinations used in the ANN models.

	Model	Input parameters
1	ANN-10	I_{sc}, V_{oc}, V_{max}, I_{max}, EFF, FF, G, v, T_{am}, T_{panel}
2	ANN-8	V_{max}, I_{max}, EFF, FF, G, v, T_{am}, T_{panel}
3	ANN-6	EFF, FF, G, v, T_{am}, T_{panel}

Source: Authors, (2026).

For each model, the number of neurons in the hidden layer was varied between 1 and 15, and the architecture was trained multiple times to identify the configuration that yielded the best pre curacy. The methodology of the ANN model is illustrated in Figure 3.

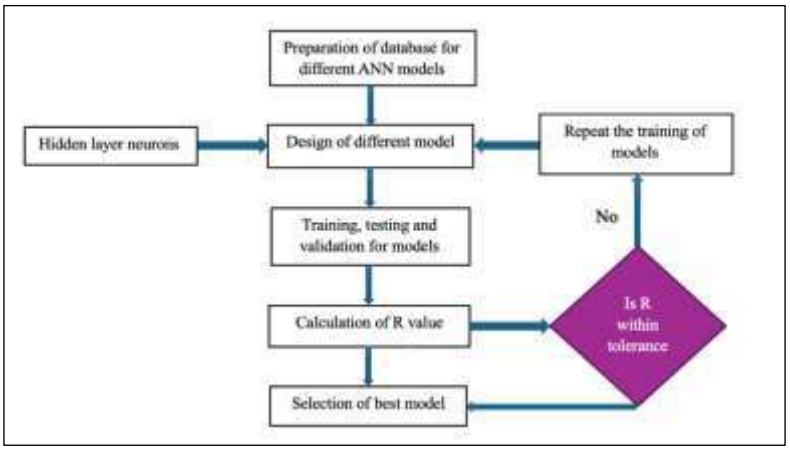


Figure 3: Implementation of methodology for ANN in the present study.
Source: Authors, (2026).

III. RESULTS

This manuscript investigates the effect of different multilayer perceptron (MLP) architectures on the performance of three artificial neural network (ANN) models for predicting solar radiation across various intensity ranges (200-1000 W/m²). The study evaluates the influence of the number of neurons in the hidden layer on the correlation coefficient (R), which measures the accuracy of model predictions. The analysis reveals that ANN-10 demonstrates consistent and high predictive performance across multiple neuron configurations, particularly excelling in the 800-1000 W/m² range Fig(4). It maintains high R-values (mostly above 0.95), reflecting stability, robustness, and optimal model complexity. ANN-8, consistently achieves high and stable R-values, indicating superior generalization and minimal oscillations across neuron configurations. In contrast, ANN-6 exhibits significant fluctuations and instability across all ranges, suggesting higher sensitivity to network structure and potential overfitting or underfitting issues. These irregular trends highlight the challenges of model complexity in this configuration.

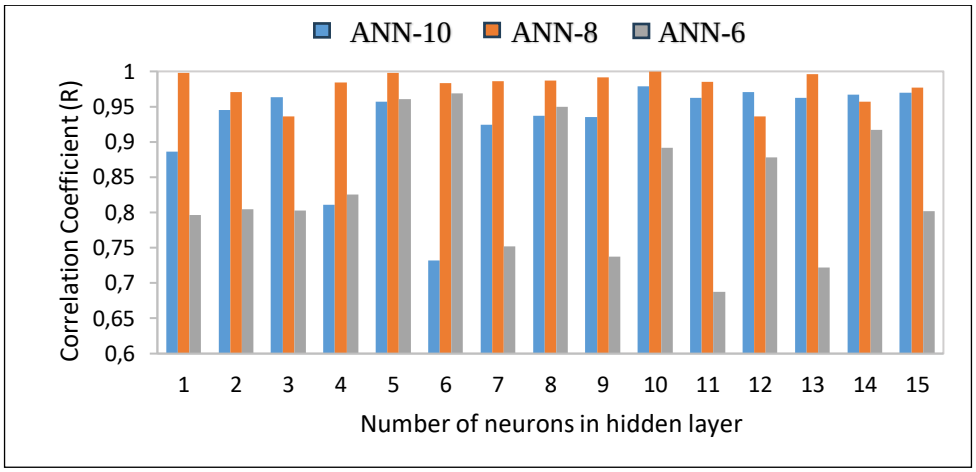


Figure 4: Effect of MLP architecture on ANN fitting performance for three models of solar radiation range (800-1000 W/m^2).
Source: Authors, (2026).

Overall, the findings suggest that while ANN-10 and ANN-8 perform well across specific ranges, ANN-10 stands out for its consistent high performance in the 800-1000 W/m² range, making it a strong candidate for solar radiation modelling in these higher intensity ranges. Figure (5) presents the effect of multilayer perceptron (MLP) architecture, specifically the number of neurons in the hidden layer, on the correlation coefficient (R) for three artificial neural network (ANN) models: ANN-10, ANN-8, and ANN-6, developed for predicting solar radiation within the range of 600-800 W/m². The correlation coefficient (R) quantifies the agreement between predicted and measured solar radiation, with values approaching unity indicating higher model accuracy and generalization capability.

As shown in the Fig (5), all three ANN configurations exhibit variations in correlation coefficient values with changes in the number of neurons in the hidden layer, reflecting the sensitivity of model performance to network complexity. The ANN-10 model (blue line) displays relatively stable and high R-values across most configurations, predominantly ranging between 0.90 and 1.00, with minor fluctuations at specific neuron counts. The ANN-8 model (orange line) also achieves strong predictive performance but exhibits more pronounced oscillations, with R-values varying between 0.80 and 1.00 depending on the number of neurons. In contrast, the ANN-6 model (green line) shows significant instability and broader fluctuations in R, ranging between 0.70 and 0.98, suggesting higher sensitivity to the chosen network structure and potential overfitting or underfitting at certain configurations.

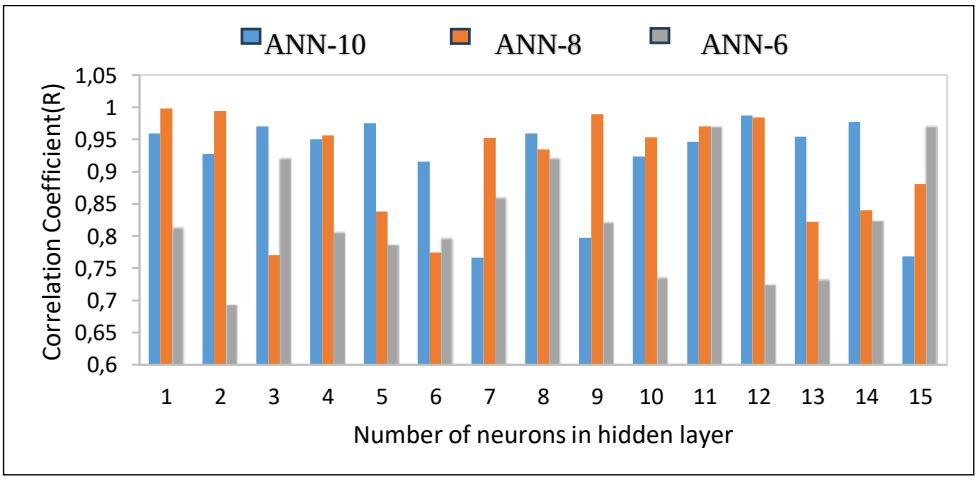


Figure 5: Effect of MLP architecture on ANN fitting for three models of solar radiation range (600-800 W/m^2).
Source: Authors, (2026).

Based on the observed R-values, the ANN-6 model demonstrates the best overall performance, maintaining high correlation coefficients (mostly above 0.95) across a wide range of hidden-layer neuron counts. This indicates superior stability, robustness, and generalization in fitting the nonlinear relationships governing solar radiation within the specified range. The consistently high R-values imply that ANN-6 achieved an optimal balance between model complexity and learning efficiency, resulting in minimized prediction errors and better adaptability to unseen data. Conversely, the irregular trends observed in ANN-8 and ANN-6 suggest less consistent learning behaviour and potential over-dependence on the number of hidden neurons.

Figure (6) presents the effect of multilayer perceptron (MLP) architecture specifically the number of neurons in the hidden layer on the correlation coefficient (R) for three artificial neural network (ANN) models: ANN-10, ANN-8, and ANN-6, developed for predicting solar radiation within the range of 400-600 W/m². The correlation coefficient (R) quantifies the agreement between predicted and measured solar radiation, with values approaching unity indicating higher model accuracy and generalization capability. As shown in the figure, all three ANN configurations exhibit variations in correlation coefficient values with changes in hidden-layer neuron numbers, reflecting the sensitivity of model performance to network complexity.

The ANN-10 model (blue line) displays relatively stable and high R-values across most configurations, predominantly ranging between 0.90 and 1.00, with minor fluctuations at specific neuron counts. The ANN-8 model (orange line) also achieves strong predictive performance but exhibits more pronounced oscillations, with R-values varying between 0.80 and 1.00 depending on the number of neurons. In contrast, the ANN-6 model (green line) shows significant instability and broader fluctuations in R, ranging between 0.70 and 0.98, suggesting higher sensitivity to the chosen network structure and potential overfitting or underfitting at certain configurations.

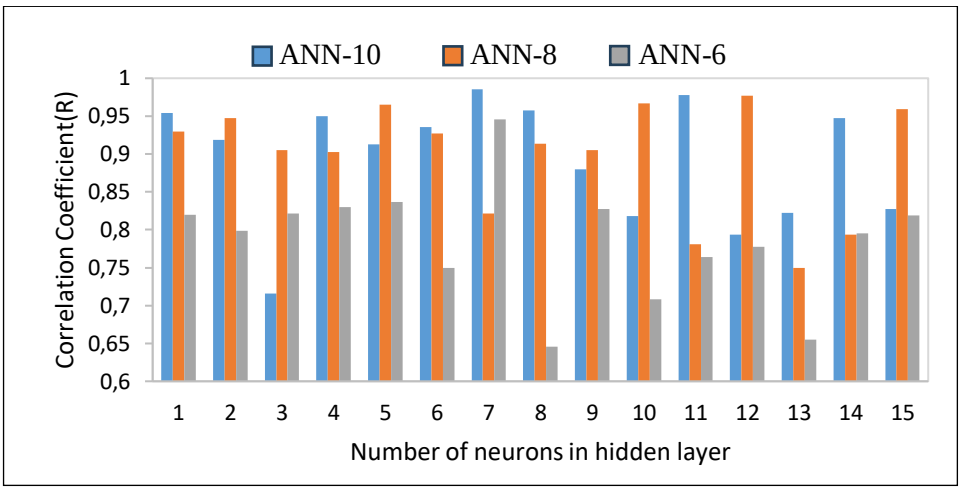


Figure 6: Effect of MLP architecture on ANN fitting for three models of solar radiation range (400-600 W/m^2).
Source: Authors, (2026).

Based on the observed R-values, the ANN-10 model demonstrates the best overall performance, maintaining high correlation coefficients (mostly above 0.95) across a wide range of hidden-layer neuron counts. This indicates superior stability, robustness, and generalization in fitting the nonlinear relationships governing solar radiation within the specified range. The consistently high R-values imply that ANN-10 achieved an optimal balance between model complexity and learning efficiency, resulting in minimized prediction errors and better adaptability to unseen data. Conversely, the irregular trends observed in ANN-8 and ANN-6 suggest less consistent learning behaviour and potential over-dependence on the number of hidden neurons.

Figure (7) presents the effect of multilayer perceptron (MLP) architecture specifically, the number of neurons in the hidden layer on the correlation coefficient (R) for three artificial neural network (ANN) models: ANN-10, ANN-8, and ANN-6, developed for predicting solar radiation within the range of 200-400 W/m². The correlation coefficient (R) represents the strength of the relationship between predicted and measured solar radiation values, where values closer to 1 indicate superior prediction accuracy and stronger model generalization. As shown in the figure, all three ANN models exhibit noticeable variations in R-values as the number of hidden-layer neurons increases, reflecting the sensitivity of model performance to architectural complexity.

The ANN-8 model (orange line) consistently demonstrates the highest and most stable R-values across nearly all configurations, maintaining correlations predominantly above 0.95, with minimal fluctuations. This trend indicates that ANN-8 effectively captures the nonlinear patterns in the solar radiation data while avoiding overfitting or underfitting. The ANN-R10 model (blue line) shows moderate performance, with R-values generally ranging between 0.80 and 0.95. Although it achieves strong correlations at certain neuron counts (e.g., around 5–6 and 10–11 neurons), the performance exhibits greater oscillations, suggesting sensitivity to the network structure and possible inconsistencies in generalization when the architecture changes. The ANN-6 model (green line), by contrast, demonstrates the lowest and most unstable R-values, fluctuating between 0.55 and 0.95, which implies weaker learning stability and a higher likelihood of overfitting or underfitting across different hidden-layer configurations.

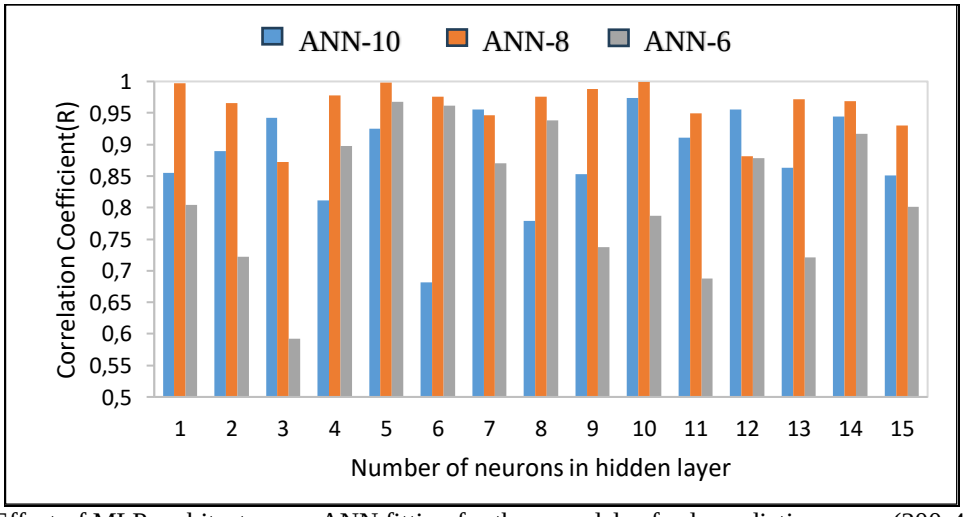


Figure 7: Effect of MLP architecture on ANN fitting for three models of solar radiation range (200-400 W/m^2).
Source: Authors, (2026).

Based on the observed results, the ANN-8 model is identified as the best-performing MLP architecture for this solar radiation range. Its consistently high R-values and minimal oscillations indicate superior predictive robustness, stability, and adaptability across varying hidden-layer neuron counts. This suggests that ANN-8 achieved an optimal balance between model complexity and learning efficiency, leading to more reliable and accurate solar radiation predictions.

VI. DISCUSSION

For manuscript discussion, these findings highlight that while all models are influenced by hidden-layer neuron variations, ANN-8's architecture demonstrates the most reliable and generalizable learning behaviour. The stability of its correlation coefficients indicates that it effectively captures the underlying physical and nonlinear dependencies of solar radiation within the 200-400 W/m² range, making it the most suitable architecture for practical modelling and forecasting applications.

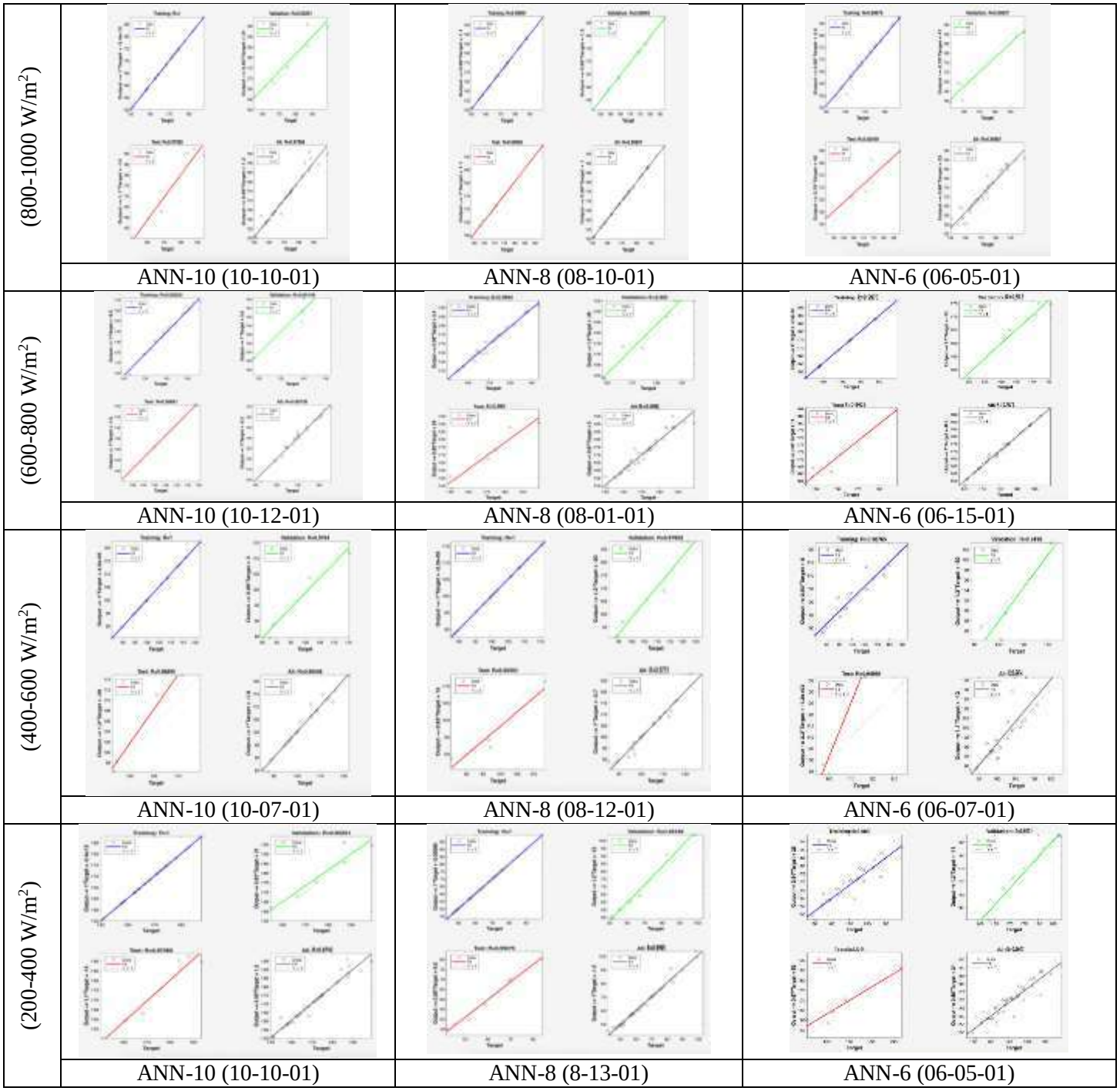


Figure 8: Regression performance of ANN-10, ANN-8, and ANN-6 models at different solar radiation ranges (200–1000 W/m²) showing correlation coefficients (R) for training, validation, testing, and overall datasets.
Source: Authors, (2026).

The regression plots for all models (ANN-10, ANN-8, and ANN-6) under different solar radiation ranges (200–1000 W/m²) are presented in Fig (8). The correlation coefficient (R) was used to evaluate the prediction accuracy of each model.

Among the tested configurations, the ANN-8 model consistently achieved the highest R-values across training, validation, and testing datasets, indicating strong predictive reliability and superior generalization capability. At higher radiation levels (800–1000 W/m²), ANN-8 achieved an almost perfect fit (R ≈ 0.9999), while ANN-10 and ANN-8 recorded slightly lower but still high correlations. Similarly, for mid-range radiation (600–800 W/m²), ANN-10 exhibited an R ≈ 0.985, demonstrating robust prediction with minimal deviation. At lower radiation levels (200–400 W/m² and 400–600 W/m²), all models maintained good predictive capability, though minor reductions in R were observed, possibly due to increased variability in measured data. Overall, ANN-8 outperformed ANN-10 and ANN-6 across all radiation categories, suggesting that its architecture and parameter tuning were most effective in capturing the nonlinear behaviour of solar radiation. The comparative evaluation of ANN architectures across different solar radiation ranges demonstrates the strong influence of input size and hidden layer configuration on model performance. The results consistently reveal that the number of input features plays a decisive role in predictive accuracy, while the hidden layer size contributes to fine-tuning the learning and generalization capability of the models [29].

For the highest radiation range (800–1000 W/m²), ANN-R8 with eight inputs emerged as the most effective configuration. The MLP (8-10-1) architecture achieved nearly perfect regression values across training, testing, and validation, highlighting its ability to capture the nonlinear relationships in this intensity band. In contrast, models with six inputs (ANN-6) showed weaker performance, with the best architecture (6-5-1) still underperforming relative to the 8- and 10-input models [30]. This suggests that reducing the dimensionality of the input space limits the ability of network to learn the complex patterns inherent in high-intensity solar radiation data. In the 600–800 W/m² range, the superiority of ANN-8 (8 inputs) was reinforced. The MLP (8-1-1) model provided the highest consistency, achieving regression values close to unity across all phases. Interestingly, in this radiation band, some 6-input models also delivered competitive performance, particularly MLP (6-15-1), which maintained strong validation accuracy. This indicates that for mid-range radiation intensities, the complexity of input features can be somewhat reduced without severely compromising prediction accuracy, provided that the hidden layer is configured effectively [31].

At the 400–600 W/m² range, the results once again emphasized the strength of the 8-input models. ANN-8 with MLP (8-15-1) delivered outstanding regression values, demonstrating robustness and generalization ability across datasets. The 10-input model (ANN-10) with MLP (10-11-1) also showed high reliability, but its performance fluctuated more across different subsets, suggesting occasional overfitting. The 6-input models in this range displayed greater variability, with certain configurations performing well (e.g., MLP 6-7-1) while others struggled, indicating that fewer input features reduce stability and predictive robustness [32]. For the lowest radiation range (200–400 W/m²), the pattern remained consistent: 8-input models were the most effective, particularly ANN-8 with MLP (8-13-1), which achieved near-perfect regression values across training, testing, and validation. The 10-input models performed well in certain configurations, such as MLP (10-7-1), which balanced accuracy and generalization effectively. However, the 6-input models again showed inconsistent performance, with only a few architectures like MLP (6-7-1) delivering reliable outcomes [33]. This reinforces the observation that reducing input dimensionality to six features constrains the capacity of model to capture variability in the lower radiation bands.

A critical aspect revealed by the results is the balance between overfitting and underfitting. Some 10-input networks, particularly those with larger hidden layers, demonstrated very high training accuracy but weaker performance on validation data. This suggests a tendency toward overfitting, where the model memorizes training patterns without effectively generalizing to unseen data [34]. On the other hand, some 6-input models, especially with fewer hidden neurons, showed underfitting, as they were unable to capture sufficient complexity to accurately represent solar radiation variability. The 8-input networks consistently avoided both extremes by maintaining strong generalization, indicating an optimal balance between model complexity and predictive stability. The findings across all radiation ranges demonstrate that ANN-8 models with eight input features consistently outperformed both the 10-input and 6-input networks. The 8-input configurations not only provided superior predictive accuracy but also ensured stability and generalization across different solar radiation ranges. The results further highlight the importance of optimizing hidden layer size, as models with too few or too many neurons occasionally exhibited reduced generalization despite high training accuracy [35].

Thus, selecting the appropriate number of inputs alongside a balanced hidden layer configuration is critical for achieving robust solar radiation prediction. Among all tested architectures, ANN-8 with MLP (8-10-1) stands out as the most effective, offering both computational efficiency and high predictive reliability across radiation ranges. From a practical perspective, these findings have direct implications for solar energy applications. Reliable solar radiation prediction is essential for photovoltaic (PV) system design, energy storage planning, and grid integration. Accurate forecasts help optimize the sizing of PV modules and batteries, improve energy dispatch strategies, and reduce reliance on fossil-fuel backup systems. The demonstrated stability of the 8-input models indicates that such configurations can provide dependable forecasts under varying radiation conditions, thereby supporting more efficient energy management. Furthermore, the insights on overfitting and underfitting emphasize the need for careful architecture selection to avoid unreliable predictions that could compromise planning and operational decisions in solar energy systems [36]. The superior performance of ANN-8 across all radiation ranges demonstrates its robustness and adaptability in modelling nonlinear solar radiation behaviour.

The near-unity correlation coefficients (R > 0.99) in both training and testing phases indicate excellent learning and generalization capability. The slightly reduced performance of ANN-10 and ANN-6 in certain ranges may be attributed to overfitting or inadequate neuron configuration, which limits the model's ability to adapt to complex variations in input–output relationships. The stability of ANN-8 across both high and low radiation levels confirms its optimized network architecture and well-balanced hidden layer design. These results affirm that ANN-based models, particularly ANN-8, are highly reliable for solar radiation estimation and serve as an efficient computational alternative to traditional empirical or statistical methods. The performance of three artificial neural network (ANN) models (ANN-10, ANN-8, and ANN-6) was evaluated across different solar radiation ranges (200–1000 W/m²) to understand the effect of multilayer perceptron (MLP) architecture, specifically the number of neurons in the hidden layer, on the correlation coefficient (R). The results revealed that ANN-10 demonstrated the most consistent and robust performance, particularly in the 800–1000 W/m² range, maintaining high R-values (between 0.90 and 1.00) across various neuron configurations, suggesting optimal stability and generalization capability.

In contrast, ANN-8 achieved the highest and most stable R-values, particularly in the 200-400 W/m² range, consistently exceeding 0.95. This indicates its superior ability to generalize and effectively capture the nonlinear relationships in the data, with minimal fluctuations in performance across hidden-layer neuron changes. However, ANN-6 exhibited significant instability across all radiation ranges, with fluctuating R-values ranging from 0.55 to 0.98, pointing to overfitting and underfitting challenges due to its sensitivity to network complexity. Overall, the findings suggest that the choice of ANN architecture plays a crucial role in predicting solar radiation, with ANN-10 being most suitable for higher radiation ranges and ANN-8 offering superior reliability in lower ranges, while ANN-6 showed less consistent results, making it less ideal for accurate predictions [37].

This section presents a comparison of current state-of-the-art regression models for photovoltaic power forecasting, evaluated using the R^2 regression metric in Table 2.

Compared to the research articles referenced in Table 1, the proposed model presented improved results in photovoltaic power prediction as measured by the R^2 regression metric. However, the novel aspect of the model is located in the portability of the most effective regression technique. The proposed model has achieved a R^2 value of 0.999 in 800–1000 W/m² range and utilises a novel preprocessing approach and method. The feature engineering process in the proposed model employed correlation coefficients to extract significant features from the dataset.

Table 2: Comparison of present results with literature.

Case study	Feature selection method	Regression model	Features Type	Maximum R	Dataset Used
[38]	Tree-based feature importance & principal component analysis	Artificial neural network and random forest	4	0.9672	Hawaii, United States of America (2016)
[39]	7, 6, and 5 input scenarios manually tested; no automatic feature selection algorithm.	(ANN) with multilayer feed-forward architecture and backpropagation algorithms (GD, LM, RP, SCG).	9	0.9629	Indian Meteorological Department
[40]	Wavelet transformation-based decomposition technique	WT-LSTM, LSTM, Ridge regression, Lasso regression, elastic-net regression	7	0.9749	Urbana Champaign, Illinois
[41]	Manual selection using domain knowledge; 6 input parameters selected for physical relevance to PV generation.	(ANN) with Levenberg–Marquardt backpropagation.	6	0.9925	Middle Technical University, Baghdad–Iraq.
[42]	Pearson's correlation and heatmap	(LR), (GBR), (MLP)	4	0.9823	Sharda University PV Dataset
Present study	Feature reduction performed only by comparing 10, 8, and 6 input combinations	Levenberg-Marquardt algorithm; ANN-10, ANN-8, ANN-6	10	0.9995	UCSI University Kuala Lumpur, Malaysia), 1-month experimental dataset (25 June–26 July 2025)

Source: Authors, (2026).

V. CONCLUSION

This study developed and evaluated ANN models for predicting the output power of photovoltaic (PV) systems under varying environmental conditions. Three ANN configurations with different input combinations (10, 8, and 6 parameters) were evaluated across four solar radiation ranges (200–400, 400–600, 600–800, and 800–1000 W/m²). The results demonstrate that model performance strongly depends on both the number of input features and the hidden layer architecture. Among the tested models, ANN-8 (8 inputs) consistently outperformed ANN-10 (10 inputs) and ANN-6 (6 inputs), achieving near-perfect regression values ($R \approx 0.999$) across most radiation ranges. This indicates that selecting an optimal subset of input variables, rather than using all available parameters, enhances predictive accuracy and reduces computational complexity. The findings also confirm that ANN models are highly effective in capturing the nonlinear relationships between electrical and environmental parameters influencing PV performance. Overall, the study highlights the robustness of ANN-based modelling for solar energy forecasting and system performance evaluation. By systematically comparing different ANN architectures, this work provides practical insights into input feature selection and network design, supporting the deployment of reliable and efficient prediction tools for PV applications across diverse climatic conditions.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Ravshanbek Rakhmatulaev, Rodney Tan, Aliev Rayimjon, Yu L. J.
Methodology: Ravshanbek Rakhmatulaev, Rodney Tan, Aliev Rayimjon, Yu L. J..
Investigation: Ravshanbek Rakhmatulaev, Rodney Tan, Aliev Rayimjon, Yu L. J..
Discussion of results: Ravshanbek Rakhmatulaev, Rodney Tan, Aliev Rayimjon, Yu L. J..
Writing – Original Draft: Ravshanbek Rakhmatulaev, Rodney Tan, Aliev Rayimjon, Yu L. J..
Writing – Review and Editing: Ravshanbek Rakhmatulaev, Rodney Tan, Aliev Rayimjon, Yu L. J..
Resources: Ravshanbek Rakhmatulaev, Rodney Tan, Aliev Rayimjon, Yu L. J..
Supervision: Ravshanbek Rakhmatulaev, Rodney Tan, Aliev Rayimjon, Yu L. J..
Approval of the final text: Ravshanbek Rakhmatulaev, Rodney Tan, Aliev Rayimjon, Yu L. J..

VII. REFERENCES

[1] O. Bashiru, C. Ochem, L. A. Enyejo, H. N. N. Manuel, and T. O. Adeoye, "The crucial role of renewable energy in achieving the sustainable development goals for cleaner energy," Global Journal of Engineering and Technology Advances, vol. 19, no. 3, pp. 11–36, 2024.

[2] M. M. Hasan et al., "Harnessing solar power: a review of photovoltaic innovations, solar thermal systems, and the dawn of energy storage solutions," Energies (Basel), vol. 16, no. 18, p. 6456, 2023.

[3] M. D. S. Hossain, A. Wadi Al-Fatlawi, L. Kumar, Y. R. Fang, and M. E. H. Assad, "Solar PV high-penetration scenario: an overview of the global PV power status and future growth," Energy Systems, pp. 1–57, 2024.

[4] P. Gupta and R. Singh, "PV power forecasting based on data-driven models: a review," International Journal of Sustainable Engineering, vol. 14, no. 6, pp. 1733–1755, 2021.

[5] K. Hasan, S. B. Yousuf, M. S. H. K. Tushar, B. K. Das, P. Das, and M. S. Islam, "Effects of different environmental and operational factors on the PV performance: A comprehensive review," Energy Sci Eng, vol. 10, no. 2, pp. 656–675, 2022.

[6] M. Amir, Zaheeruddin, and A. Haque, "Intelligent based hybrid renewable energy resources forecasting and real time power demand management system for resilient energy systems," Sci Prog, vol. 105, no. 4, p. 00368504221132144, 2022.

[7] M. R. Vogt et al., "Introducing a comprehensive physics-based modelling framework for tandem and other PV systems," Solar Energy Materials and Solar Cells, vol. 247, p. 111944, 2022.

[8] D. Saadaoui et al., "Parameters extraction of single diode and double diode models using analytical and numerical approach: A comparative study," International Journal of Modelling and Simulation, vol. 45, no. 2, pp. 569–598, 2025.

[9] G. M. Tina, C. Ventura, S. Ferlito, and S. De Vito, "A state-of-art-review on machine-learning based methods for PV," Applied Sciences, vol. 11, no. 16, p. 7550, 2021.

[10] S. Sivamani, M. Udayakumar, and N. Sellappan, "Prediction of single cylinder direct injection diesel engine performance fuelled with lemon peel oil biodiesel using artificial neural network," Mater Today Proc, vol. 80, pp. 456–463, 2023.

[11] Z. Zhang et al., "Spot Market Electricity Price Forecast via the Combination of Transformer and Ornstein-Uhlenbeck Process," in 2025 21st International Conference on the European Energy Market (EEM), IEEE, 2025, pp. 1–6.

[12] A. P. Selvam and S. N. S. Al-Humairi, "Environmental impact evaluation using smart real-time weather monitoring systems: a systematic review," Innovative Infrastructure Solutions, vol. 10, no. 1, pp. 1–24, 2025.

[13] M. Jaber, A. S. Abd Hamid, K. Sopian, A. Fazlizan, and A. Ibrahim, "Prediction model for the performance of different PV modules using artificial neural networks," Applied Sciences, vol. 12, no. 7, p. 3349, 2022.

[14] J. H. Yousif and H. A. Kazem, "Prediction and evaluation of photovoltaic-thermal energy systems production using artificial neural network and experimental dataset," Case Studies in Thermal Engineering, vol. 27, p. 101297, 2021.

[15] R. B. Roy et al., "A comparative performance analysis of ANN algorithms for MPPT energy harvesting in solar PV system," IEEE Access, vol. 9, pp. 102137–102152, 2021.

[16] R. El Chaal and M. O. Aboutafail, "A comparative study of back-propagation algorithms: Levenberg-Marquart and BFGS for the formation of multilayer neural networks for estimation of fluoride," Commun. Math. Biol. Neurosci., vol. 2022, p. Article-ID, 2022.

[17] R. García-Ródenas, L. J. Linares, and J. A. López-Gómez, "Memetic algorithms for training feedforward neural networks: an approach based on gravitational search algorithm," Neural Comput Appl, vol. 33, no. 7, pp. 2561–2588, 2021.

[18] E. Pashaei and E. Pashaei, "Training feedforward neural network using enhanced black hole algorithm: a case study on COVID-19 related ACE2 gene expression classification," Arab J Sci Eng, vol. 46, no. 4, pp. 3807–3828, 2021.

[19] Q. Askari and I. Younas, "Political optimizer based feedforward neural network for classification and function approximation," Neural Process Lett, vol. 53, no. 1, pp. 429–458, 2021.

[20] A. Singh and K. Gaurav, "Deep learning and data fusion to estimate surface soil moisture from multi-sensor satellite images," Sci Rep, vol. 13, no. 1, p. 2251, 2023.

[21] A. Alsirhani, M. M. Alshahrani, A. Abukwaik, A. I. Taloba, R. M. Abd El-Aziz, and M. Salem, "A novel approach to predicting the stability of the smart grid utilizing MLP-ELM technique," Alexandria Engineering Journal, vol. 74, pp. 495–508, 2023.

[22] M. B. A. Shuvho, M. A. Chowdhury, S. Ahmed, and M. A. Kashem, "Prediction of solar irradiation and performance evaluation of grid connected solar 80KWp PV plant in Bangladesh," Energy Reports, vol. 5, pp. 714–722, 2019.

[23] A. Mellit, S. Sağlam, and S. A. Kalogirou, “Artificial neural network-based model for estimating the produced power of a photovoltaic module,” Renew Energy, vol. 60, pp. 71–78, 2013.

[24] S. A. Rizzo and G. Scelba, “ANN based MPPT method for rapidly variable shading conditions,” Appl Energy, vol. 145, pp. 124–132, 2015.

[25] R. May, G. Dandy, and H. Maier, “Review of input variable selection methods for artificial neural networks,” Artificial neural networks-methodological advances and biomedical applications, vol. 10, no. 1, pp. 19–45, 2011.

[26] M. M. Bejani and M. Ghatee, “A systematic review on overfitting control in shallow and deep neural networks,” Artif Intell Rev, vol. 54, no. 8, pp. 6391–6438, 2021.

[27] S. Sivamani, S. Selvakumar, K. Rajendran, and S. Muthusamy, “Artificial neural network–genetic algorithm-based optimization of biodiesel production from Simarouba glauca,” Biofuels, vol. 10, no. 3, pp. 393–401, 2019.

[28] P. S. Kumar and S. Sivamani, “Numerical analysis and implementation of artificial neural network algorithm for nonlinear function,” International Journal of Information Technology, vol. 13, no. 5, pp. 2059–2068, 2021.

[29] M. A. Ali, A. Elsayed, I. Elkabani, M. Akrami, M. E. Youssef, and G. E. Hassan, “Optimizing artificial neural networks for the accurate prediction of global solar radiation: A performance comparison with conventional methods,” Energies (Basel), vol. 16, no. 17, p. 6165, 2023.

[30] M. Mukhtar et al., “Development and comparison of two novel hybrid neural network models for hourly solar radiation prediction,” Applied Sciences, vol. 12, no. 3, p. 1435, 2022.

[31] S. Ghimire et al., “Hybrid convolutional neural network-multilayer perceptron model for solar radiation prediction,” Cognit Comput, vol. 15, no. 2, pp. 645–671, 2023.

[32] A. A. Imam, A. Abusorrah, M. M. A. Seedahmed, and M. Marzband, “Accurate forecasting of global horizontal irradiance in Saudi Arabia: a comparative study of machine learning predictive models and feature selection techniques,” Mathematics, vol. 12, no. 16, p. 2600, 2024.

[33] S. Y. Heng et al., “Artificial neural network model with different backpropagation algorithms and meteorological data for solar radiation prediction,” Sci Rep, vol. 12, no. 1, p. 10457, 2022.

[34]bB. Girdhani and M. Agrawal, “Performance evaluation of statistical and deep learning models for daily solar global horizontal radiation prediction: implications for renewable energy and sustainability,” Environ Dev Sustain, pp. 1–29, 2025.

[35] A. K. Yadav, R. Kumar, M. Wang, G. Fekete, and T. Singh, “Comparative analysis of daily global solar radiation prediction using deep learning models inputted with stochastic variables,” Sci Rep, vol. 15, no. 1, p. 10786, 2025.

[36] G. Drałus, D. Mazur, J. Kusznier, and J. Drałus, “Application of artificial intelligence algorithms in multilayer perceptron and Elman networks to predict photovoltaic power plant generation,” Energies (Basel), vol. 16, no. 18, p. 6697, 2023.

[37] V. Moya-Almeida, B. Diezma-Iglesias, E. Correa-Hernando, C. Vaquero-Miguel, and N. Alvarado-Arias, “Setpoint temperature estimation to achieve target solvent concentrations in S. cerevisiae fermentations using inverse neural networks and fuzzy logic,” Eng Appl Artif Intell, vol. 127, p. 107248, 2024.

[38] U. Munawar and Z. Wang, “A Framework of Using Machine Learning Approaches for Short-Term Solar Power Forecasting,” Journal of Electrical Engineering and Technology, vol. 15, no. 2, pp. 561–569, Mar. 2020, doi: 10.1007/S42835-020-00346-4.

[39] N. Premalatha and A. Valan Arasu, “Prediction of solar radiation for solar systems by using ANN models with different back propagation algorithms,” Journal of Applied Research and Technology, vol. 14, no. 3, pp. 206–214, Jun. 2016, doi: 10.1016/J.JART.2016.05.001.

[40] M. Mishra, P. Byomakesha Dash, J. Nayak, B. Naik, and S. Kumar Swain, “Deep learning and wavelet transform integrated approach for short-term solar PV power prediction,” Measurement (Lond), vol. 166, Dec. 2020, doi: 10.1016/J.MEASUREMENT.2020.108250.

[41] A. T. Mohammad, H. M. Hussen, and H. J. Akeiber, “Prediction of the output power of photovoltaic module using artificial neural networks model with optimizing the neurons number,” International Journal of Renewable Energy Development, vol. 12, no. 3, pp. 478–487, May 2023, doi: 10.14710/IJRED.2023.49972.

[42] B. U. D. Abdullah, S. A. Khanday, N. U. Islam, S. Lata, H. Fatima, and S. H. Nengroo, “Comparative Analysis Using Multiple Regression Models for Forecasting Photovoltaic Power Generation,” Energies 2024, Vol. 17, Page 1564, vol. 17, no. 7, p. 1564, Mar. 2024, doi: 10.3390/EN17071564.